Privacy Preservation in Multi-Cloud Secure Data Fusion for Infectious-Disease Analysis

Jianqing Liu, *Member, IEEE* Chi Zhang, *Member, IEEE* Kaiping Xue, *Senior Member, IEEE* Yuguang Fang, *Fellow, IEEE*

Abstract—It is often observed that people's data are scattered across various organizations and these data can be used to generate usable insights when integrated. However, data fusion from multiple data hosting sites could put user privacy at risk albeit with some security mechanisms. This paper studies a data-analytic platform that adopts the Kulldorff scan statistic to determine infectious-disease spatial hotspots by integrating and analyzing users' health and location data that are respectively stored in two clouds. We examine the privacy threats to this platform which has a key-oblivious inner product encryption (KOIPE) mechanism in place to ensure that only coarse-grained statistical data is revealed to the honest-but-curious (HbC) entity. To protect user privacy from the designed inference attack, we exploit a game-theoretic approach to incentivize users to form anonymous clusters with a quantitative privacy guarantee. We conduct extensive simulations based on real-life datasets to demonstrate the performance of our scheme in terms of design overhead and privacy level.

Index Terms—public health, Kulldorff scan statistic, secure multi-party computation, Bayesian inference, game theory.

1 Introduction

Classified as one of the top leading causes of death in the United States, infectious disease is a serious public health problem [2]. The impact of infectious diseases is immense but unfortunately, rapid urbanization and globalization increases the vulnerability of our society to its outbreaks. This year's COVID-19 pandemic is a typical but brutal example of how threatening infectious disease could be. Therefore, efficient detection and timely response to infectious disease outbreaks, should it occur, are the key steps for public health organizations. Among many analysis interests, the spatial clustering analysis is of critical importance [3]. By analyzing people's health (e.g., fever, coughing) and location (e.g., zipcode) data, epidemiologists could identify geographical disease clusters at the early stage of the disease outbreak. Then, public resources like antibiotic prophylaxis could be allocated to prevent its further dissemination. Recent deployment of disease monitoring systems has gained great attention. COVID-19 web dashboard was launched to track global COVID cases [4]. Another project called Biological-Agent Correlation Tracker (BACTracker) [5] deployed by MIT Lincoln Laboratory aims to mitigate possible bioterrorist attacks.

However, these participatory-based mobile systems bear the weakness such as the coarse timeliness, limited representativeness and unreliable participation rate. For example, some systems collect data weekly; a majority of participants are women; and participation rates also relate to illness with first-time participants being more likely to be sick than repeated ones. Furthermore, it has also been noted that patients are generally very reluctant to report their health and location information for a variety reasons, some related to socio-demographic differences and others for privacy concerns (e.g., unwanted intrusive marketing [6]). Ideally, high-fidelity clustering analysis requires as much an individual's information as possible, but existing mobile systems fail in this regard.

To remedy this problem, the popularity of cloud services may shed a light on an alternative solution. Nowadays, our digital life are scattered among a myriad places in the cloud — our location data at Google, health data at Apple and social network at Facebook. Ideally, integrating these data from multiple clouds will create many insightful knowledge about the public health [7], but some hurdles such as rigid business models, intellectual property (IP) concerns, legal and ethical issues challenge the practicality of this multi-cloud model. Amongst these obstacles, privacy is the utmost concern to users so many security mechanisms are widely developed to protect user's privacy in multi-cloud data fusion. For example, in our preliminary work [1], we designed a secure multi-party computation (SMC)-based scheme which only gives statistical data to the honest-but-curious (HbC) data-integration entity for the analysis of spatial clusters of infectious diseases. At first glance, disclosing statistical data — a common methodology in the state-of-the-art — can preserve user's privacy because an adversary only has a fuzzy view of a group, but this syntactic privacy

Part of this work was previously presented in IEEE Symposium of Privacy-Aware Computing [1].

J. Liu is with the Department of Electrical and Computer Engineering, The University of Alabama in Huntsville (UAH), Huntsville, AL 35899 USA e-mail: jianqing.liu@uah.edu

C. Zhang and K. Xue are with the department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China email: {chizhang,kpxue}@ustc.edu.cn

Y. Fang is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA e-mail: fang@ece.ufl.edu

The work by J. Liu was supported in part by National Science Foundation under grant ECCS-2047484 and the New Faculty Research Award by UAH. The work by C. Zhang supported in part by the Natural Science Foundation of China (NSFC) under Grants 61871362 and U19B2023. The work by K. Xue supported in part by the Youth Innovation Promotion Association of the Chinese Academy of Sciences (CAS) under Grant No. Y202093. The work by Y. Fang was supported in part by National Science Foundation under grant IIS-1722791.

model is vulnerable to the inference attack by the adversary with prior knowledge, which are recently evidenced in a few works [8], [9], [10], [11].

In this paper, departing from our prior work [1], we investigate the privacy implication and countermeasure in a secure multi-cloud data fusion model for infectious disease analysis. Specifically, this work makes the following major contributions.

- We present a novel framework to collect users' multi-institutional data across various cloud platforms for spatial clustering analysis of infectious diseases.
- We develop a SMC-based scheme by leveraging the key-oblivious inner product encryption (KOIPE) mechanism to ensure that untrusted entities can only get statistical data of a group instead of an individual.
- We demonstrate the effectiveness of Bayesian inference attack on statistical data in deriving an individual's data. Then, we propose a non-perturbative game-theoretic approach to preserve user privacy and guarantee high-fidelity data analysis by incentivizing participation in this multi-cloud platform.

The remainder of the paper is organized as follows. Section 2 surveys the related works. Section 3 describes the system model and security assumptions. Section 4 introduces the preliminaries for our design. Section 5 presents the SMC-based data fusion protocol. Then we discuss an effective inference attack in Section 6 and propose its countermeasure in Section 7. The performance evaluation of our scheme is later shown in Section 8. Finally, we conclude the paper in Section 9.

2 Related Works

2.1 Infectious disease monitoring and analysis

Spatial analysis of infectious diseases has drawn great attention recently. One line of research efforts is to develop surveillance platforms for data collection. Exemplary platforms include the COVID-19 web dashboard [4]. However, these systems have unreliable participation rate and the collected data has limited representativeness.

With the popularity of cloud services like social networks, an alternative yet effective solution becomes available for infectious disease monitoring. For instance, Twitter data is exploited to track flu [12] and Ebola [13]. Fung *et al.* showed the effectiveness of using Weibo data to track 42 infectious diseases like Dengue and Malaria in China [14]. Zhang *et al.* proposed a privacy-preserving framework for integrating health cloud with social cloud to predict users' health condition based on their social contacts [15].

2.2 Privacy and security in infectious disease analysis

To ensure privacy when data is collected, distributed and consumed in health applications, randomization (e.g., differential privacy [16]) and anonymization are common techniques. However, these approaches tend to either introduce large distortion leading to errors in the data analysis, or suffer from re-identification attacks, nullifying the effort of protection practices [17].

Another idea on preserving privacy is to reduce the unnecessary data disclosure as much as possible, preferably only revealing statistical information. Rmind [18] is a cryptographic tool that provides secure computation for statistical analysis. It implements a number of secure operations including quantiles, co-variance, statistical tests, to name a few. However, it has been shown in recent works that privacy is not preserved in a platform that only reveals statistical data through secure computations [8], [9]. A simple inference algorithm like random forest would reveal users' information in an aggregated statistical dataset, which is coined as the membership inference attack [19]. As the countermeasure, PrivaDA [10] and Pettai *et al.* [11] introduced differential privacy to an aggregated dataset to preserve user privacy.

In order to secure data analytics from multiple sources, SMC has been widely adopted recently. For example, Laud *et al.* [20] proposed an SMC protocol to remove duplicated records among multiple databases to avoid record linkage. However, Ah-Fat *et al.* [11] recently observed that some information of sensitive input could still be leaked from SMC outputs.

In this work, we intend to address a missing element in existing works and use a non-perturbative technique to protect user privacy from inference attack on statistical information, especially the one from the SMC output. Besides privacy, such a technique shall not hurdle infectious disease monitoring system from collecting high-fidelity data. In other words, it should incentivize users to contribute their data for analysis.

3 System Model

In this section, we give a high-level discussion on the system model for data fusion, and the security model.

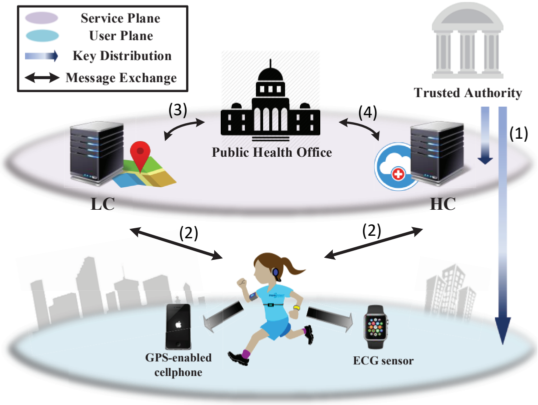


Figure 1: System model for the data fusion framework exploiting two cloud services.

3.1 System Model

As shown in Fig.1, our system consists of five entities. Collectively, they are the trusted authority (TA), public health office (PHO), location-based service (LBS) cloud server (LC), health service cloud server (HC) and users.

Users get services from LC and HC through off-the-shelf devices such as their GPS-enabled mobile phones and wearable devices. LC and HC are operated by different enterprises and they offer location and health services to users, respectively. PHO is a public health entity that conducts disease monitoring and analysis. Upon users' consent, PHO can collect data from LC and HC. TA bootstraps the system, generates and distributes secret materials to other entities. It also addresses disputes and revokes misbehaved entities if needed. The rationale of assuming three independent and inter-operable entities is based on some practical systems and business examples. For instance, amid COVID-19 pandemic, Apple and Google developed respective APIs in their app stores to collect users' location (i.e., physical contact) and health information in order to deliver contract warnings and enable health authorities (e.g., CDC) to keep track of the spread of COVID-19 [21]. The developed schemes in this paper could potentially be integrated in the existing platforms for privacy protection.

3.2 Security Model

TA is fully trusted by other entities in the system and it cannot be compromised by the adversary. LC and HC are honest-but-curious (HbC). That is to say, they honestly follow the protocol but are curious about users' location and health information. Their incentives for malicious behaviors include delivering commercials, denial of insurance for unhealthy users and many more. PHO is assumed HbC as well in the sense that it honestly conducts statistical analysis of infectious disease but are curious of one's location and health data for purposes like segregating infected patients, which however is against users' willingness and thus compromising their privacy. PHO, LC and HC are operated by separate entities and they are assumed to be non-colluding. Users in this system are not trusted. They may launch Sybil attacks to mislead PHO's analysis by injecting fake/duplicated data to LC and HC. Their incentive is either to cause panic in an uninfected region or to reduce PHO's awareness of an infected area. This applies to bio-terrorists or businesses trying to gain commercial advantages over others.

4 Preliminaries

4.1 Kulldorff Scan Statistic [22]

The Kulldorff scan statistic was firstly proposed in 1997 [22], and it now becomes a powerful tool in performing both spatial and temporal clustering analysis for infectious diseases. In spatial analysis, the Kulldorff scan statistic is able to discover small regions (e.g., a school or a shopping mall) of significantly elevated disease density. In what follows, we describe how the Kulldorff scan statistic works.

A surveillance region $\mathcal{K}$ is divided into subareas $\{s_1, s_2, ..., s_K\}$ of any arbitrary level of resolution, and $\mathcal{K} = \bigcup_{i=1}^{K} s_i$. The disease headcount and population in each subarea, denoted as $\{c_1, c_2, ..., c_K\}$ and $\{p_1, p_2, ..., p_K\}$, respectively, are collected. Let the total disease case count $C_{tot} = \sum_{i=1}^{K} c_i$ and census population $P_{tot} = \sum_{i=1}^{K} p_i$ of whole region $\mathcal{K}$. The Kulldorff scan statistic is then applied to search all possible clusters of adjacent subareas for abnormal ones with disease overdensity. Specifically, suppose $\{S_1, S_2, ..., S_M\}$ is the set of such clusters, each of which has disease case count C_j and population P_j. The Kulldorff spatial scan statistic proceeds to calculate the respective cluster density D_j as

$$C_j \log \frac{C_j}{P_j} + (C_{tot} - C_j) \log \frac{C_{tot} - C_j}{P_{tot} - P_j} - C_{tot} \log \frac{C_{tot}}{P_{tot}}, \quad (1)$$

if $\frac{C_j}{P_j} > \frac{C_{tot}}{P_{tot}}$ and 0, otherwise.

In so doing, the maximum density $mrd = \max_{S_j} D_j$ and the corresponding cluster $mdr = \arg\max_{S_j} D_j$ in the region $\mathcal{K}$ can be identified. To evaluate if this cluster is statistically significant, the Kulldorff spatial scan statistic assumes that c_i follows inhomogeneous Poisson processes and a randomization testing approach is conducted to examine mdr. The statistical significance (i.e., p-value) is then calculated so that the cluster is considered as the outlier or being statistically significant when $p \leq 0.05$.

4.2 Privacy Metric

We follow the popular privacy metric [8] and model user privacy level as the inverse of an adversary's capability in correctly inferring a user's private information. Specifically, *Area Under Curve (AUC)* is used to capture the adversary's overall inference performance while the *privacy loss (PL)* score is calculated to represent a particular user's loss of privacy.

AUC Score: Suppose that a user's private information is $x \in \{0, 1\}$ and adversary's inference is $x^* \in \{0, 1\}$. We have the following metrics:

- True Positive (TP) when $x^* = 1$ and $x = 1$;
- True Negative (TN) when $x^* = 0$ and $x = 0$;
- False Positive (FP) when $x^* = 1$ and $x = 0$;
- False Negative (FN) when $x^* = 0$ and $x = 1$;

from which we could derive the True Positive Rate as TPR = TP/(TP+FN) and False Positive Rate as FPR = FP/(FP+TN). Then, based on different discrimination thresholds, the Receiver Operating Characteristic (ROC) curve could be obtained and the AUC is just the area under the ROC curve which captures the adversary's overall performance in inferring the user's information x.

PL Score: Suppose the ideal case where the adversary has no prior knowledge about a user's private information and thus it has to randomly guess (AUC = 0.5) x^*. Through the inference attack, AUC may increase and we define the PL score as the adversary's relative belief gain over its original random guess:

$$PL = \begin{cases} \frac{AUC - 0.5}{0.5} & \text{if } AUC > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

5 Secure Multi-Cloud Data Fusion for Infectious Disease Analysis

5.1 Protocol Overview

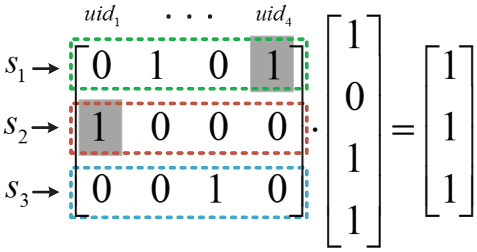
The overall information flow is shown in Fig.1. At system bootstrap, TA generates and distributes security materials to PHO, LC, HC and users, and then it can go offline (coined as flow (1) in Fig.1). Whenever an infectious disease outbreaks, PHO first determines an incentive to motivate users to contribute their data, as will be discussed in Section 7. LC and HC collect location and health data from users, respectively, coined as flow (2) in Fig.1, in which an authentication scheme based on *anonymous group signatures* is designed to avoid Sybil attacks and an anonymity scheme based on *identity-based encryption* is designed to provide controllable linkability of users' data on HC and LC. Finally, PHO interacts with LC and HC to obtain users' aggregated health and location data, coined as flow (3) and (4) in Fig.1, for spatial analysis.

For the sake of brevity, we will only present the design for flow (3) and (4), i.e., the SMC-based design for statistical data calculation, to motivate our study on the privacy threat and protection in Section 6. For more details on our design for flow (1) and (2), we refer readers to our preliminary work [1].

5.2 Secure Data Aggregation via SMC

Since the Kulldorff scan statistic only requires the census data such as the population P_j and the count of infected users C_j in each geographical grid j, the design goal is to limit PHO to access only these statistical (i.e., aggregated) data. Moreover, spatial analysis deals with prohibitively large datasets, so high efficiency is desired in the computation. Specifically, the steps of our design are as follows.

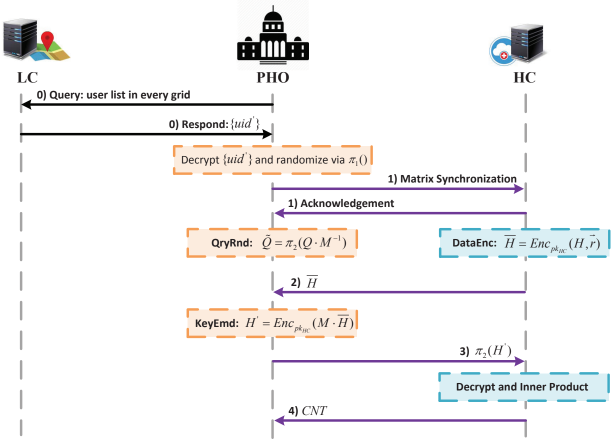
PHO starts with matching users' health and location records respectively at HC and LC by linking users' encrypted identifiers $\boldsymbol{uid}'$ at LC and $\boldsymbol{uid}$ at HC[1]. After that, PHO and HC jointly calculate the aggregated disease counts in a secure manner. For demonstrative purposes, we show a toy example in Fig.2. On the one hand, PHO forms a query matrix $\boldsymbol{Q}$ containing users' existence in one geographical grid. A value 1 represents the user in that grid; 0 otherwise. On the other hand, HC maintains users' infected status in vector $\boldsymbol{H}$. Then, PHO can efficiently calculate the disease count vector $\boldsymbol{CNT}$ in each geographical grid via the inner product of $\boldsymbol{Q}$ and $\boldsymbol{H}$. In our design, PHO's query matrix $\boldsymbol{Q}$ should be hidden to HC (for location privacy) and conversely HC's health vector $\boldsymbol{H}$ should be kept private to PHO (for health privacy). Therefore, our scheme is boiled down to a secure multiparty computation (SMC) design where PHO and HC collaboratively calculate the inner product of $\boldsymbol{Q}$ and $\boldsymbol{H}$ without revealing further information.



uid_1
. . .
uid_4
s_1
s_2
s_3
$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

Figure 2: A toy example for the batch query: PHO's query matrix contains 3 grids and 4 users; HC holds 4 users' infected status; the inner product gives the number of infected users in each grid.

Inspired by the secure k nearest neighbour (kNN) scheme [23], which is detailed in Appendix A, we modify its original design because (i) it only provides relative distance instead of exact value; (ii) the random matrix is known to both entities, which cannot satisfy the security requirement for our design. In light of this, we propose a Key-Oblivious Inner Product Encryption (**KOIPE**) scheme to address these problems. The overview of our design is shown in Fig.3 and the detailed description is given as follows.



LC
PHO
HC
0) Query: user list in every grid
0) Respond: $\{uid'\}$
Decrypt $\{uid'\}$ and randomize via $\pi_1()$
1) Matrix Synchronization
1) Acknowledgement
QryRnd: $\tilde{Q} = \pi_2(Q \cdot M^{-1})$
DataEnc: $\overline{H} = Enc_{pk_{HC}}(H, \vec{r})$
2) $\overline{H}$
KeyEmd: $H' = Enc_{pk_{HC}}(M \cdot \overline{H})$
3) $\pi_2(H')$
Decrypt and Inner Product
4) CNT

Figure 3: Diagram for information exchange and secure data aggregation.

(1) **MtxSync**: PHO and HC first "synchronize" their matrices so that mismatched records will be eliminated. To do it securely, PHO applies a permutation mechanism $\boldsymbol{\pi_1}$ to randomize the user list $\boldsymbol{uid}$, which is then sent to HC to trim and re-order $\boldsymbol{H}$.

(2) **QryRnd**: PHO selects a random invertible matrix $\boldsymbol{M}$ of size $N \times N$ to encrypt the query matrix into $\overline{\boldsymbol{Q}} = \boldsymbol{Q} \cdot \boldsymbol{M}^{-1}$. To enhance security, another permutation vector $\boldsymbol{\pi_2}$ is applied to $\overline{\boldsymbol{Q}}$ and PHO then sends the transformed encrypted query matrix $\overline{\overline{\boldsymbol{Q}}}$ (i.e., $\overline{\overline{\boldsymbol{Q}}} = \boldsymbol{\pi_2}(\overline{\boldsymbol{Q}})$) to HC. Note that applying the same permutation on two matrices, regardless of the number of times, does not change the inner product of them.

(3) **DataEnc**: We use the additive homomorphic property from the Paillier cryptosystem [24]. HC firstly obtains a key pair (pk_{HC}, sk_{HC}) from TA. Then, HC encrypts $\boldsymbol{H}$ using public key pk_{HC} and gets $\overline{\boldsymbol{H}}$, which is sent back to PHO. Specifically, for $h_i \in \boldsymbol{H}$, the encryption runs as $\overline{h_i} = Enc_{pk_{HC}}(h_i, r_i)$ where $Enc_{pk_{HC}}()$ is the encryption function and r_i is a random number selected in correspondence with h_i.

(4) **KeyEmd**: Here, we intend to securely embed the random invertible matrix $\boldsymbol{M}$ in $\overline{\boldsymbol{H}}$. By leveraging the property of the additive Paillier cryptosystem, PHO calculates $\boldsymbol{H}' = Enc_{pk_{HC}}(\boldsymbol{M} \cdot \overline{\boldsymbol{H}})$ via the following arithmetic computa-

[1] Linkability is ensured by our design in flow (2) as shown in Fig.1.

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/TMC.2022.3145745, IEEE Transactions on Mobile Computing

tion for each element in $\overline{\boldsymbol{H}}$:

$$\begin{aligned} h_i' &= \prod_{j=1}^{N} Enc_{pk_{HC}}(\overline{h_j}, r_j)^{m_{i,j}} \\ &= \prod_{j=1}^{N} Enc_{pk_{HC}}(m_{i,j}\overline{h_j}, r_j^{m_{i,j}}) \\ &= Enc_{pk_{HC}}(\sum_{j=1}^{N} m_{i,j}\overline{h_j}, \prod_{j=1}^{N} r_j^{m_{i,j}}), \quad 1 \le i \le N. \end{aligned}$$

Then, PHO applies the same permutation $\boldsymbol{\pi_2}$ as in Step 2 to randomize the $\boldsymbol{H}'$, and then sends it along with the encrypted query matrix $\overline{\overline{\boldsymbol{Q}}}$ back to HC.

(5) **InPrd**: Upon receiving the two matrices from PHO in Step 4, HC first decrypts the health data vector using the secret key sk_{HC} into $\overline{\overline{\boldsymbol{H}}}$ and then computes the inner product to derive the number of infected users as $\boldsymbol{CNT} = \overline{\overline{\boldsymbol{Q}}} \cdot \overline{\overline{\boldsymbol{H^T}}}$. Then, HC sends it back to PHO, which concludes the whole process. With the disease case count c_j and the respective number of participants p_j for every geographical grid area s_j, PHO can apply the Kulldorff scan statistic to search for the spatial clusters that exhibit statistical significance.

The above design can be proven to be secure in the sense that neither HC nor PHO can determine a specific user's health and location data purely based on the interactions in Fig.3. The reasoning is as follows. First, by observing PHO's $\overline{\overline{Q}}$, HC has to solve a system of linear equations which has K equations but $K \cdot N$ variables to derive the randomization matrix M so as to revert Q. Although secure kNN scheme is inherently vulnerable to known-plaintext attack (KPA) [23] implying that HC can invert M with sufficient number of plaintext-ciphertext pairs, KPA is practically rare especially given our threat model that PHO and HC will not collude. Therefore, there lacks sufficient information to find M which implies that HC cannot determine a user's location data. Second, PHO is oblivious of a user's location due to the batch query to the LC; and it cannot deduce a user's health status either since solving discrete logarithm problem is assumed to be mathematically hard. Nonetheless, PHO can access the final output statistics, i.e., how many users are in a grid area and how many of them carry the disease. This statistics could then be misused by the PHO to infer a user's sensitive information. To illustrate this point, we showcase PHO's inference capability in the following section.

6 Inference Attack on Statistical Data

In this section, we evaluate how capable PHO is in compromising individual's health data depending on its side information and inference capability. Consider a scenario where PHO monitors the city-level public health in real-time. That is to say, PHO continuously collects data from HC and LC to facilitate its statistical analysis. By correlating data from different time instants, PHO could increase its belief of a particular user's health condition, which on the contrary reduces the user's privacy level. Next, we demonstrate the effectiveness of such an inference attack based on a real-world dataset. We then propose a game-theoretic approach in Section 7 as the countermeasure.

6.1 Adversary's Knowledge and Objective

Following the same notations from previous sections, we use $\boldsymbol{Q}$ to denote the locations of all participant users, where $q_{i,j} \in \boldsymbol{Q}^{|K| \times |N|}$ is 1 if user i is within geographical grid area j and 0 otherwise. By adding the vector of disease counts $\boldsymbol{CNT}$ into $\boldsymbol{Q}$, we construct the matrix $\boldsymbol{D} \in \mathbb{Z}^{|K| \times (|N|+1)}$ as PHO's knowledge after querying LC and HC. Suppose PHO periodically collects $\boldsymbol{D}$ across a finite series $\{1, 2, ..., t\}$ for the purpose of real-time disease surveillance. Then, by time t, PHO has a stream of observed data denoted as $\boldsymbol{S_t} = \{\boldsymbol{D_1}, \boldsymbol{D_2}, ..., \boldsymbol{D_t}\}$ and we let $\boldsymbol{S_t}[i] = \boldsymbol{D_i}$ for $1 \le i \le t$.

Next, we assume that PHO possesses certain prior knowledge $\mathcal{P}$ regarding how likely users could get infectious disease. For instance, certain infectious diseases (e.g., airborne or foodborne [25]) exhibit evident age and gender bias [26]. PHO could easily construct $\mathcal{P}$ by correlating users' wellness with their physiological attributes which are potentially revealed from the mobility pattern in $\boldsymbol{S_t}$.

In addition, suppose PHO has an inference function *adv* which takes (u^*, $\boldsymbol{S_t}$, $\mathcal{P}$) as input and yields an inferred result about whether user u^* carries infectious disease or not. Mathematically, the inference attack can be characterized as

$$h^* = adv(u^*, \boldsymbol{S_t}, \mathcal{P}),$$

where $h^* \in \{0, 1\}$ represents the inferred result. Note that although PHO's inference is w.r.t. a user with pseudonym uid^*, the potential re-identification attacks, e.g., from mobility traces, could undermine our faith on the anonymization system [27].

6.2 Bayesian Inference Attack

Here, we opt to instantiate PHO's inference function *adv* with the *Bayesian inference model* [28]. Given the prior knowledge $\mathcal{P}$, PHO derives the posterior knowledge when observing the collected dataset $\boldsymbol{D}$, which is coined as *evidence*. This process repeats for t times, which captures PHO's up-to-date belief in users' health condition. Specifically, by time t, PHO's belief in whether user i has the disease or not is characterized by Eq.(2) and Eq.(3), respectively. Here, the health condition $\boldsymbol{H}$ is a set of random variables, which are assumed independent with each other and follow the non-identical Bernoulli distributions.

In Eq.(2), the equality (a) is for the calculation of Bayesian posterior probability via the chain rule where $\boldsymbol{S_t}$ is the evidence; $h_i = 1$ is the hypothesis that user i carries the disease; and $\Pr(h_i = 1)$ is part of PHO's prior knowledge $\mathcal{P}$ regarding user i's health condition. Recall that every $\boldsymbol{D}$ consists of users' location information $\boldsymbol{Q}$ and the count of infected users in each geographic grid k. Suppose the monitoring phase t is at the early stage of disease developments, so the infected users only exhibit the signs and symptoms but the disease is not that contagious [29]. We thus assume no users are infected by the infected ones through their social contacts within time t. As a result, the event $\boldsymbol{D_i}$ is independent of any other event $\boldsymbol{D_j}$ and the conditional probability can be re-written followed by the equality (b). Furthermore, define a set $\mathcal{L}_{i|t} = \{u_j \in \mathcal{U} | q_{s,i|t} \wedge q_{s,j|t} = 1, \exists s \in \mathcal{K}, 1 \le i \ne j \le N\}$ as the set of users who are at the same geographic grid s with user i at a given time stamp t, and the operand ∧ is the Boolean **AND** operation. Then, the number of infected users where the user i resides at time stamp t is $\sum_{\{k | u_k \in \mathcal{L}_{i|t}, k \ne i, 1 \le k \le N\}} h_k = c_{i|t}$, which is collected from HC. Suppose users' mobility pattern $\boldsymbol{Q}$ is an observed

$$\begin{aligned}\Pr(h_i=1|\boldsymbol{S_t}) &\overset{(a)}{=} \frac{\Pr(\boldsymbol{D_t}|h_i=1,\boldsymbol{S_{t-1}})\cdots\Pr(\boldsymbol{D_1}|h_i=1)\cdot\Pr(h_i=1)}{\Pr(\boldsymbol{S_t})} \overset{(b)}{=} \frac{\Pr(\boldsymbol{D_t}|h_i=1)\cdots\Pr(\boldsymbol{D_1}|h_i=1)\cdot\Pr(h_i=1)}{\prod_{j=1}^{t}\Pr(\boldsymbol{D_j})}\\ &\overset{(c)}{=} \frac{\Pr(\sum\limits_{\{k|u_k\in\mathcal{L}_{i|t},1\le k\le N\}} h_k=c_{i|t}|h_i=1)\cdots\Pr(\sum\limits_{\{k|u_k\in\mathcal{L}_{i|1},1\le k\le N\}} h_k=c_{i|1}|h_i=1)\cdot\Pr(h_i=1)}{\prod_{j=1}^{t}\Pr(\sum\limits_{\{k|u_k\in\mathcal{L}_{i|j},1\le k\le N\}} h_k=c_{i|j})}\\ &\overset{(d)}{=} \frac{\Pr(\sum\limits_{\{k|u_k\in\mathcal{L}_{i|t},k\neq i,1\le k\le N\}} h_k=c_{i|t}-1)\cdots\Pr(\sum\limits_{\{k|u_k\in\mathcal{L}_{i|1},k\neq i,1\le k\le N\}} h_k=c_{i|1}-1)\cdot\Pr(h_i=1)}{\prod_{j=1}^{t}\Pr(\sum\limits_{\{k|u_k\in\mathcal{L}_{i|j},1\le k\le N\}} h_k=c_{i|j})}, 1\le i\le N. \qquad (2)\end{aligned}$$

$$\Pr(h_i=0|\boldsymbol{S_t}) = \frac{\Pr(\sum\limits_{\{k|u_k\in\mathcal{L}_{i|t},k\neq i,1\le k\le N\}} h_k=c_{i|t})\cdots\Pr(\sum\limits_{\{k|u_k\in\mathcal{L}_{i|1},k\neq i,1\le k\le N\}} h_k=c_{i|1})\cdot\Pr(h_i=0)}{\prod_{j=1}^{t}\Pr(\sum\limits_{\{k|u_k\in\mathcal{L}_{i|j},1\le k\le N\}} h_k=c_{i|j})}, 1\le i\le N. \qquad (3)$$

and thus deterministic variable whereas the occurrence of an event $\boldsymbol{D_t}$ is the outcome of the set of random variables in $\boldsymbol{H}$. Therefore, we have the equality (c). Recall that users' health conditions are independent with each other, so we can further write the conditional probability according to the equality (d) and derive the final expression for the Bayesian posterior probability as in Eq.(2). Similarly, PHO's belief in user i's health condition as non-infected can be captured by Eq.(3) or simply as $\Pr(h_i = 0|\boldsymbol{S_t})$ = 1 - $\Pr(h_i = 1|\boldsymbol{S_t})$.

Since the set of random variables in $\boldsymbol{H}$ are independent and follow non-identical Bernoulli distributions, the sum of the subset of $\boldsymbol{H}$ (i.e., $\sum\limits_{k} h_k$) follows the Poisson binomial distribution and we could calculate the Bayesian posterior probabilities Eq.(2) and Eq.(3) in this regard. Then, we select a threshold ζ such that user i is classified to have the disease (i.e., $h_i^* = 1$) if $\Pr(h_i = 1|\boldsymbol{S_t}) \ge \zeta$ while not having the disease (i.e., $h_i^* = 0$) otherwise. For various selections of ζ, we can construct the ROC curve for a specific user and then use AUC and PL as the metric to quantify PHO's inference capability and user's privacy loss, respectively.

6.3 Evaluation

6.3.1 Simulation Setup

We choose a real-world dataset that captures the mobility characteristics, obtained from Gowalla [30], a location-based social networking website where users share their location by checking-ins. Specifically, it contains a total of 6,442,892 check-ins of users over the period of February 2009 to October 2010. Each record consists of a user identifier, time stamp, latitude, longitude and location id. For our analysis, we focus on the users that have checked in on March 14, 2010 in Austin, Texas, USA. The geographical area of Austin was approximated by taking a rectangular grid with corners at coordinates (29.5°N, 98.5°W), (29.0°N, 97.5°W), (30.5°N, 97.5°W) and (30.5°N, 98.5°W); and center at coordinate (30.0°N, 98.0°W). A total of 10,638 check-ins were reported by 1,801 unique users in Austin within the specified time.

To determine PHO's inference capability, we perform a series of inference attack for each user. First, the region of interest is split into grids (e.g., 10×10, 15×15 and 20×20). Next, the infected rate of population is assumed to be 10% and remain constant over the considered time and PHO's prior knowledge on each user's disease condition is set to 0.5. We then randomly assign disease condition (i.e., 0 or 1) to every user based on the infected rate. Next, suppose PHO collects the statistical data in a time granularity of hours (e.g., 1 hr., 2 hrs. and 3hrs) within a day. Given these inputs, PHO updates its posterior probability using Eq.(2) and Eq.(3) for each user. To address the randomness, we repeat the above steps for 100 iterations and calculate the average value.

6.3.2 Results and Analysis

Fig.4 plots cumulative distribution function (CDF) versus the AUC score achieved by the Bayesian inference model under different inference settings. We observe that decreasing the size of grid area (i.e., increasing the number of grids) or increasing data collection granularity/frequency results in higher mean AUC scores. For instance, the mean AUC score for the setting of 10×10 grid size and time granularity 1 hour is 0.72. While for the same grid size but time granularity of 2 hours, the mean AUC score is 0.66.

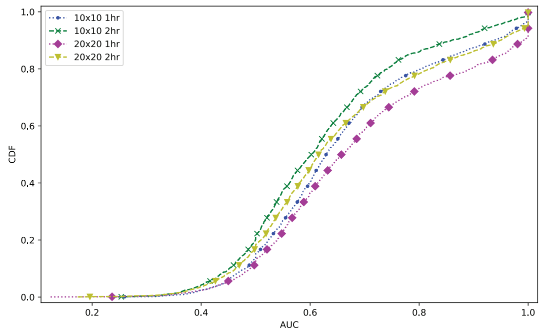


Figure 4: PHO's inference capability under different inference settings.

We also plot PL over different inference settings as shown in Fig.5. This indicates the relative belief gain of the PHO over its initial random guess which reflects the impact of the Bayesian inference on the privacy loss of each user. We observe that for a certain infected rate, PL increases as the size of grid decreases or as the time granularity increases. For example, 20×20 grid size and 1-hour time granularity

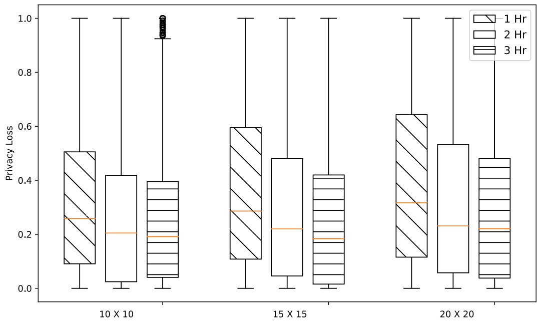


Figure 5: Privacy loss per-user under different inference settings.

gives the mean PL score 0.31; but for the same grid size and time granularity 3-hours, the mean PL score is 0.22.

7 Countermeasure to Inference Attack

The take-away from above simulations is that small aggregation group size (or anonymity set) and continuous statistical observation are the major obstacles to the privacy preservation. Thus, we intend to develop a mechanism to protect user's privacy from the aforementioned statistical inference attack. To the best of our knowledge, existing solutions include creating a larger aggregation area (e.g., by zipcode) [31], obfuscating individual or statistical data (e.g., location and health data) [32], changing users' pseudonyms frequently at *mix zone* [27], [33], etc. However, these approaches have their inherent disadvantages: the large aggregation spatial region could compromise the sensitivity of disease analysis [34]; and updating pseudonyms is burdensome for mobile users which may discourage them from participation [35]. In this section, we propose a game-theoretic approach to incentivize participation and also naturally create a large anonymity group so that every user's data is *"hidden in the crowd"*.

7.1 Game-theoretic Approach

As illustrated earlier in Section 1, a general volunteer-based data collection system could be limited by the participatory rate and data representiveness. Our system bears the similar issue if user privacy is not properly preserved. Therefore, we will explore *how much incentive PHO should provide to a user so that it would compensate user privacy loss and thus motivating her to contribute the data*. By proper design, PHO could obtain sufficient data for infectious disease analysis while preserving the user's data privacy.

We formulate the problem by resorting to the *Stackelberg Bayesian game* model [36], where the self-interested users (i.e., the followers) determine their optimal strategy (participate or not) based on the incentive offered by PHO (i.e., the leader). Firstly, we characterize each user's utility function as

$$\mathcal{U}_i = s_i \gamma P - \theta_i \frac{s_i}{\sum_{j=1}^{N} s_j},$$

where user i's potential strategy $s_i \in \{0, 1\}$ indicates her decision on contributing data ($s_i = 1$) or not ($s_i = 0$). P is the amount of incentive paid by PHO to each participatory user, and we apply consistent P without customizing it for different users. γ is a system parameter unifying user's received incentive and its privacy level. Recall that a user's privacy level is related to adversary's inference capability. Without assuming the awareness of such inference capability, user's privacy level can be simply characterized by using the $k-$anonymity concept, which is $\frac{s_i}{\sum_{j=1}^{N} s_j}$. Note that in practise where adversaries have side information, $k-$anonymity could be sacrificed so our defined privacy level serves an upper-bound for different adversarial settings. θ_i captures user i's *type* which characterizes her valuation (or weight) of privacy. Note that one user could be oblivious of others' type and thus the utility function. Hence, we are dealing with the incomplete information scenario and the game played among users is a *Bayesian Game*. To address the uncertainty, we follow the classical work by Harsanyi [37] where users are only aware of the distribution $\mathcal{F}$ from which user's type θ_i is sampled. Besides, user i's utility function could be further re-written as follows due to the discrete nature of s_i:

$$\mathcal{U}_i(s_i, \mathcal{S}_{-i}; \theta_i, \Theta_{-i}) = \begin{cases} 0, & s_i = 0 \\ \gamma P - \theta_i \frac{1}{1+\sum_{j\neq i} s_j}, & s_i = 1 \end{cases} \tag{4}$$

On the other hand, by incentivizing users to contribute their data, PHO receives the following payoff:

$$\mathcal{U}_0 = \lambda \log(1 + \sum_{j=1}^{N} s_i) - P \sum_{j=1}^{N} s_i, \tag{5}$$

where λ is a system parameter like γ and the logarithmic function reflects PHO's diminishing return on the number of participating users for its Kulldorff scan statistic analysis.

Under this game-theoretic model, the objective of PHO is to find the optimal value of P to maximize Eq.(5), while each user's goal is to find her optimal strategy s_i to maximize Eq.(4) given other users' strategies $\mathcal{S}_{-i}$ and the incentive P. This model falls into the category of *Stackelberg Bayesian game* where PHO is the game *leader* initiating a payment P and users are the game *followers* responding their participating strategies given P. In what follows, we start with the game of followers and determine users' best responses as a function of the payment P. We then find the optimal incentive strategy P^* of PHO based on the result of followers' game to maximize Eq.(5).

Due to the incomplete information game of followers, we have the following definition of the equilibrium state:

Definition 7.1 (Bayesian Nash Equilibrium (BNE))**.** *Given the user's belief about the types of other users θ_i and others' strategies $\mathcal{S}_{-i}$, a strategy profile $\mathcal{S}^* = \{s_1^*, s_2^*, ..., s_N^*\}$ is a BNE if s_i^* for every user i maximizes her expected utility. That is to say,*

$$s_i^*(\theta_i) = \arg\max_{s_i} \sum_{\Theta_{-i}} f(\theta_{-i}) \times \mathcal{U}_i(s_i, \mathcal{S}_{-i}^*; \theta_i, \theta_{-i}) \tag{6}$$

Since users' types are sampled from the same distribution $\mathcal{F}$ and the user's utility is a non-increasing function of type, the best response can be determined in a threshold structure [35], [38] as follows:

$$s_i^*(\theta_i) = \begin{cases} 0, & \theta_i > t^* \\ 0/1, & \theta_i = t^* \\ 1, & \theta_i < t^* \end{cases} \tag{7}$$

where t^* is the type gap between two actions 0 and 1. Thus, determining the user's optimal strategy $s_i^*(\theta_i)$ is equivalent to finding t^*. Interestingly, we see that the user of type t^* has no preference over action 0 or 1 so we have the following theorem to determine the strategy threshold.

Theorem 7.1. *Given the incentive P offered by PHO, the optimal strategy profile of users at the Bayesian Nash Equilibrium is in Eq.(7), where the strategy threshold is as follows:*

$$t^* = \begin{cases} 1-(1-\gamma\cdot N\cdot P)^{\frac{1}{N}}, & \text{if } P<\frac{1}{\gamma N} \\ 1, & \text{otherwise} \end{cases} \tag{8}$$

Proof. See Appendix B. ■

This finding concludes the solution for the game of followers and our next attempt is to determine the optimal incentive strategy P^* for PHO given users' responses in Eq.(7). As the types of users are unknown to PHO, instead of maximizing the utility function in Eq.(5), PHO should find P^* that maximizes its *expected payoff*. Based on the type distribution $\mathcal{F}$ and Eq.(7), we know that the probability of action 1 being taken is $F(t^*)$, which is the cumulative distribution function (CDF) evaluated at t^*. If we let $z=\sum_{j=1}^N s_j^*$ be the event indicating the number of participating users, z follows a binomial distribution with success/true probability $F(t^*)$ and we denote it as $\mathcal{G}$. Hence, PHO's *expected payoff* function can be written as follows:

$$\overline{\mathcal{U}_0} = \sum_z \left[\lambda\log(1+z) - P\cdot z\right]\times g(z), \tag{9}$$

where $g(z)$ is the probability mass function (PMF) of $\mathcal{G}$ evaluated at z. The PHO's attempt is to find the optimal incentive strategy, i.e., $P^* = \arg\max_P \overline{\mathcal{U}_0}$ where $P\in[0,\infty)$.

Unfortunately, Eq.(9) is in discrete nature due to z and when the number of users gets large, the closed form of Eq.(9) is infeasible to track which results in the inefficiency in solving the optimal value P^*. In light of this, we transform Eq.(9) by first obtaining the *expected* number of participating users. This heuristic will lead PHO's objective into the following form:

$$\overline{\mathcal{U}'_0} = \lambda\log[1+N\cdot F(t^*)] - P\cdot N\cdot F(t^*), \tag{10}$$

where t^* is obtained from Eq.(8) and PHO needs to find the optimal incentive such that $P^*=\arg\max_P \overline{\mathcal{U}'_0}$. Then, we have the following result which will facilitate for the search of P^*.

Theorem 7.2. *There exists a unique Stackelberg Equilibrium $(P^*,\mathcal{S}^*)$ in the leader's game, where P^* is the unique maximizer for Eq.(10) over $P\in[0,\infty)$.*

Proof. Assume that user type θ is uniformly distributed over $[0,1]$. Then, by taking Eq.(8) into Eq.$\overline{(10)}$, we can easily calculate the second order derivative of $\overline{\mathcal{U}'_0}$ as

$$\frac{d^2\overline{\mathcal{U}'_0}}{dP^2} = -\lambda\frac{N^3X^{-\frac{2N-2}{N}} - N(N-1)X^{-\frac{2N-1}{N}}Y}{Y^2} - P\cdot N(N-1)X^{-\frac{2N-1}{N}} < 0$$

where $X=(1-\gamma\cdot N\cdot P)$ and $Y=\left(1+N-N\cdot X^{\frac{1}{N}}\right)$. Therefore, the transformed payoff function $\overline{\mathcal{U}'_0}$ is strictly concave with respect to (w.r.t.) P for $P\in[0,\infty)$, albeit not smooth at $P=\frac{1}{\gamma N}$. Besides, $\overline{\mathcal{U}'_0}$ is 0 when $P=0$; while $\overline{\mathcal{U}'_0}$ is $-\infty$ as P goes to $+\infty$ because of $F(t^*)=1$ for $P=+\infty$. There exists a unique maximizer P^* that can be efficiently computed by bisection or Newton's method [39]. ■

8 Performance Evaluation

In this section, we numerically evaluate the incurred computation overhead for HC and PHO during the SMC-based data fusion phase.[2] Moreover, we examine how the game-theoretic approach could motivate users' participation and thus improve their privacy against Bayesian inference attack.

8.1 Simulation Setup

We use a workstation with 3.4GHz Intel(R) Core(TM) i7 CPU and 32GB memory to emulate running environment for users, LC/HC and PHO. We employ 2048-bit modulus as the secret key length, as with RSA, in the Paillier cryptosystem. The implementation is based on the open-source platform by John Bethencourt [40] which was built upon the GNU Multiple Precision Arithmetic Library (GMP) in C language.

We exploit one dataset which is the cancer incidence at New York State [41] for Kulldorf scan statistics. It was constructed by collecting 67,217 tumor incidences from 2005-2009 out of an average of 19.34 million population (2010 population census) covering 13,848 spatial groups. Given this dataset, we only filter the lung tumor incidence and apply the random-drop scheme to adjust the number of users, the disease count and the number of clusters, which are control variables to examine the performance of our designed scheme.

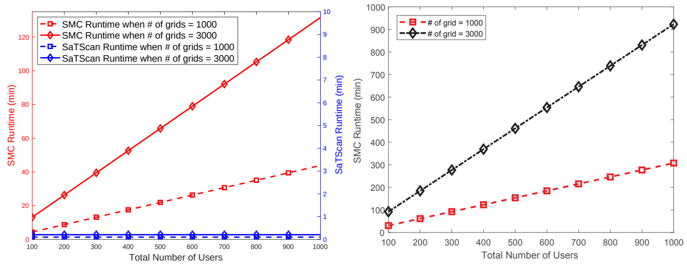
Furthermore, we explore how to steer users' behaviors under different incentives to "hide" users' private data in a large sample group. Suppose there are $N=200$ users inside one geographical grid. The system parameters γ and λ are set as control variables that represent user's bias between the incentive and privacy and the PHO's preference of utility over payment, respectively.

8.2 Computation Overhead Analysis

In our preliminary work [1], we illustrated the theoretical element-wise analysis on the computation overhead for every protocol step, so we omit it here and focus on examining numerically the performance of the SMC-based data fusion design.

Specifically, we demonstrate an end-to-end computation overhead, which consists of running SMC protocol to obtain the statistical data and using Kulldorff scan statistic to find the spatial clusters. In so doing, we aim to compare how much additional overhead is incurred due to the non-functional security design. On the one hand, we implement the KOIPE-based SMC protocol based on John Bethencourt's Paillier platform [40]. On the other hand, we use the simulator provided by the open-source SaTScan as in [42] to perform Kulldorff scan statistic. The setting is that

[2] The results of communication overhead were given in our preliminary work [1] thus are omitted here.



(a) PHO

(b) HC

Figure 6: Computation overhead for PHO and HC.

inhomogeneous Poisson process is used to test whether an area of disease over-density is indeed statistically significant among others. In every statistical analysis, we generate $R = 999$ replications and set p to 0.05. Both SMC and SaTScan simulations are run for 10 independent times to remove randomness.

The results is shown in Fig.6, and key observations from this simulation are three-fold. (1) The SMC runtime is larger than SaTScan runtime and such effect becomes more evident as we have more users and finer grids. (2) The SaTScan runtime is only dependent on the granularity of spatial grids, but irrelevant to the number of users. (3) HC's runtime is much larger than PHO's. For our first observation, the incurred overhead due to the SMC protocol is acceptable (in minute-scale) considering that a city-level disease monitoring system may only collect data in hourly or daily basis. For the second observation, this is due to the nature of Kulldorff scan statistic which analyzes spatial clusters while the disease and user counts are normalized according to Eq.1. For the third observation, the reason is that PHO only runs arithmetic modular exponentiation/multiplication while HC computes both Paillier's encryption and decryption functions.

8.3 Privacy and Incentive under Game-theoretic Approach

Since γ and λ are system parameters, their selections can greatly impact the solution to the game-theoretical model. In Fig.7, we demonstrate how the optimal payment/incentive P changes w.r.t. γ and λ. Obviously, PHO inclines to offer higher P as γ decreases for any λ. This is because users lean towards privacy preservation, especially when $\gamma < 1$, and as a result, PHO has to offer higher incentive P to maximize its utility. Another interesting observation is that as λ increases for any γ, PHO's incentive increases exponentially and then remains constant after a threshold level. This threshold implies the minimum P that invites all users to participate.

Based on above numerical guideline for parameter selection, next we examine how the PHO's incentive affects users' decision and in turn impacts PHO's own payoff. In this evaluation, we set $\gamma = 0.005$ and $\lambda = 80$, implying that user prefers privacy to incentive while PHO prefers utility to payment. Fig.8 shows the results for users' strategy and PHO's payoff w.r.t. the incentive provided by PHO. We observe that the threshold t^* increases slowly at first and steeply at the end as the PHO's payment increases, which implies that most users choose not to participate for a small incentive (otherwise, privacy will be sacrificed), but all of them decide to join simultaneously when the payment can compensate their privacy loss and a large anonymity set is formed. Besides, for $P \geq 1$, $t^* = 1$ always holds. Clearly, the consensus in participation is users' best strategy as it provides them the largest aggregation group with the minimum privacy loss. On the other hand, PHO's payoff is maximized when all users participate, but it drops as payment continues to increase which is due to an obvious reason — utility remains the same while overall payment increases as shown in Eq.(10). Therefore, the optimal strategy for PHO is to offer a minimum payment, here $P = 1$, that can invite all users to participate.

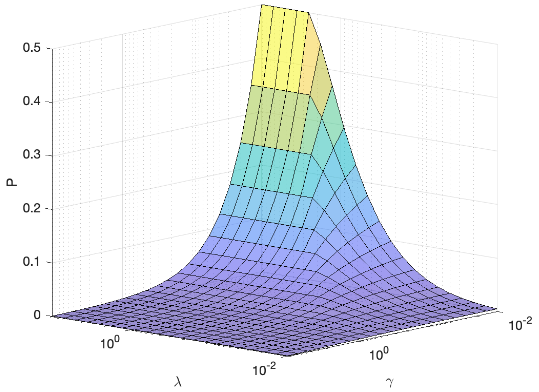


Figure 7: Selection and Impact of system parameters.

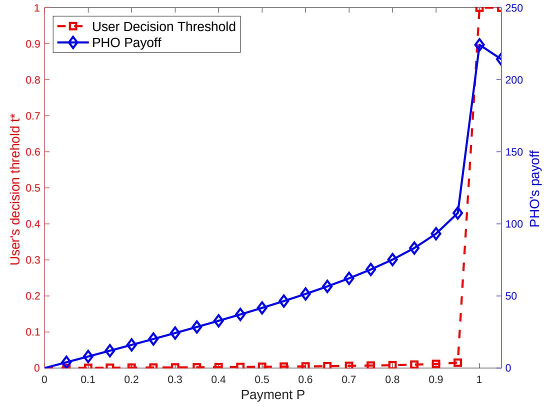


Figure 8: Users' and PHO's strategy under the game model.

Guided by the optimal strategy obtained from the game theoretic approach, Fig.9 shows how the minimum requirement for the aggregation group size could impact on user privacy. The simulation is run based on the same setup as mentioned in Section 6.3.1. The infection rate is set to 10%, the grid size to 10 × 10, and the inference frequency to 1 hr. It can be observed that privacy loss decreases when more users contribute to the aggregation statistics. For instance, for a small aggregation size of 20 users, the average privacy loss can be reduced by as large as 28% (from 0.24 in Fig.5 to 0.18 in Fig.9). In addition, this strategy is shown to benefit the outlier users whose PL was 1 when no privacy-preservation was in place as shown in Fig.5. Nevertheless, most users' privacy cannot be perfectly protected (i.e., privacy loss = 0) because the game-theoretic model, as a strategy-making model for agents of conflicting interests, in general falls short of offering provable privacy. One may seek stronger privacy concepts such as differential privacy [43] to remedy it.

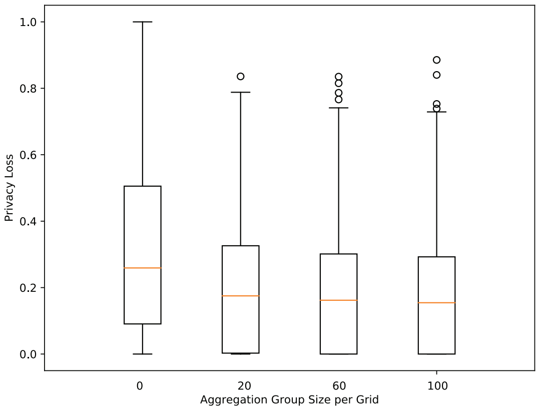


Figure 9: The impact of the minimum aggregation group size on user privacy loss.

9 Conclusion

In this paper, we have examined the privacy threat to a multi-cloud secure data fusion model for infectious-disease analysis. By designing a simple yet effective Bayesian inference technique, we have shown its impact on user's privacy loss due to the reveal of just statistical data from a developed secure multi-party computation protocol. To preserve privacy, a game-theoretic approach is proposed to defend against Bayesian inference attacks and incentivize rational users to contribute their data for public health. Numerical simulations based on real-life dataset are conducted and have quantitatively demonstrated the design overhead and privacy gain.

Acknowledgement

Authors would thank anonymous editor(s) and reviewers for their comments on our work. A special thanks is given to Dr. Apostolos Pyrgelis and Mr. Sopan Sarkar for their suggestions and insights on our proposed Bayesian inference attack.

References

[1] J. Liu, Y. Hu, H. Yue, Y. Gong, and Y. Fang, "A cloud-based secure and privacy-preserving clustering analysis of infectious disease," in *2018 IEEE Symposium on Privacy-Aware Computing (PAC)*. IEEE, 2018, pp. 107–116.

[2] H. Nichols, "The top 10 leading causes of death in the united states," 2017. [Online]. Available: https://www.medicalnewstoday.com/articles/282929.php

[3] J. Cordes and M. C. Castro, "Spatial analysis of covid-19 clusters and contextual factors in new york city," *Spatial and Spatio-temporal Epidemiology*, vol. 34, p. 100355, 2020.

[4] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," *The Lancet infectious diseases*, vol. 20, no. 5, pp. 533–534, 2020.

[5] A. Szpiro, B. Johnson, and D. Buckeridge, "Health surveillance and diagnosis for mitigating a bioterror attack," *Lincoln Laboratory Journal*, vol. 17, no. 1, 2007.

[6] K. El Emam, J. Hu, J. Mercer, L. Peyton, M. Kantarcioglu, B. Malin, D. Buckeridge, S. Samet, and C. Earle, "A secure protocol for protecting the identity of providers when disclosing data for disease surveillance," *Journal of the American Medical Informatics Association*, vol. 18, no. 3, pp. 212–217, 2011.

[7] R. Dautov, S. Distefano, and R. Buyya, "Hierarchical data fusion for smart healthcare," *Journal of Big Data*, vol. 6, no. 1, p. 19, 2019.

[8] A. Pyrgelis, C. Troncoso, and E. De Cristofaro, "Knock knock, who's there? membership inference on aggregate location data," *arXiv preprint arXiv:1708.06145*, 2017.

[9] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, "Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 1241–1250.

[10] F. Eigner, A. Kate, M. Maffei, F. Pampaloni, and I. Pryvalov, "Differentially private data aggregation with optimal utility," in *Proceedings of the 30th Annual Computer Security Applications Conference*. ACM, 2014, pp. 316–325.

[11] M. Pettai and P. Laud, "Combining differential privacy and secure multiparty computation," in *Proceedings of the 31st Annual Computer Security Applications Conference*. ACM, 2015, pp. 421–430.

[12] K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using twitter data: demonstration on flu and cancer," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1474–1477.

[13] M. Odlum and S. Yoon, "What can we learn about the ebola outbreak from tweets?" *American journal of infection control*, vol. 43, no. 6, pp. 563–571, 2015.

[14] I. C.-H. Fung, Y. Hao, J. Cai, Y. Ying, B. J. Schaible, C. M. Yu, Z. T. H. Tse, and K.-W. Fu, "Chinese social media reaction to information about 42 notifiable infectious diseases," *PLoS One*, vol. 10, no. 5, p. e0126092, 2015.

[15] K. Zhang, X. Liang, J. Ni, K. Yang, and X. Shen, "Exploiting social network to enhance human-to-human infection analysis without privacy leakage," *IEEE Transactions on Dependable and Secure Computing*, 2016.

[16] J. Liu, C. Zhang, and Y. Fang, "Epic: A differential privacy framework to defend smart homes against internet traffic analysis," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1206–1217, 2018.

[17] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 2008, pp. 111–125.

[18] D. Bogdanov, L. Kamm, S. Laur, and V. Sokk, "Rmind: a tool for cryptographically secure statistical analysis," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 3, pp. 481–495, 2016.

[19] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Towards demystifying membership inference attacks," *arXiv preprint arXiv:1807.09173*, 2018.

[20] P. Laud and A. Pankova, "Privacy-preserving record linkage in large databases using secure multiparty computation," *BMC medical genomics*, vol. 11, no. 4, pp. 33–46, 2018.

[21] Apple, "Apple and google partner on covid-19 contact tracing technology," 2020. [Online]. Available: https://www.apple.com/newsroom/2020/04/apple-and-google-partner-on-covid-19-contact-tracing-technology/

[22] M. Kulldorff, "A spatial scan statistic," *Communications in Statistics-Theory and methods*, vol. 26, no. 6, pp. 1481–1496, 1997.

[23] W. K. Wong, D. W.-l. Cheung, B. Kao, and N. Mamoulis, "Secure knn computation on encrypted databases," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 2009, pp. 139–152.

[24] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 1999, pp. 223–238.

[25] M. E. Craft, "Infectious disease transmission and contact networks in wildlife and livestock," *Phil. Trans. R. Soc. B*, vol. 370, no. 1669, p. 20140107, 2015.

[26] L. Yang, K. H. Chan, L. K. Suen, K. P. Chan, X. Wang, P. Cao, D. He, J. M. Peiris, and C. M. Wong, "Impact of the 2009 h1n1 pandemic

on age-specific epidemic curves of other respiratory viruses: A comparison of pre-pandemic, pandemic and post-pandemic periods in a subtropical city," *PloS one*, vol. 10, no. 4, p. e0125447, 2015.

[27] 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*. IEEE, 2012, pp. 972–980.

[28] G. E. Box and G. C. Tiao, *Bayesian inference in statistical analysis*. John Wiley & Sons, 2011, vol. 40.

[29] C. Fraser, S. Riley, R. M. Anderson, and N. M. Ferguson, "Factors that make an infectious disease outbreak controllable," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 16, pp. 6146–6151, 2004.

[30] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1082–1090.

[31] N. H. Fefferman, E. A. O'Neil, and E. N. Naumova, "Confidentiality and confidence: is data aggregation a means to achieve both?" *Journal of public health policy*, vol. 26, no. 4, pp. 430–449, 2005.

[32] S. C. Wieland, C. A. Cassa, K. D. Mandl, and B. Berger, "Revealing the spatial distribution of a disease while preserving privacy," *Proceedings of the National Academy of Sciences*, pp. pnas–0 801 021 105, 2008.

[33] J. Freudiger, M. H. Manshaei, J.-P. Hubaux, and D. C. Parkes, "On non-cooperative location privacy: a game-theoretic analysis," in *Proceedings of the 16th ACM conference on Computer and communications security*. ACM, 2009, pp. 324–337.

[34] K. L. Olson, S. J. Grannis, and K. D. Mandl, "Privacy protection versus cluster detection in spatial epidemiology," *American Journal of Public Health*, vol. 96, no. 11, pp. 2002–2008, 2006.

[35] X. Liu, K. Liu, L. Guo, X. Li, and Y. Fang, "A game-theoretic approach for achieving k-anonymity in location based services," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 2985–2993.

[36] R. Gibbons, *A primer in game theory*. Harvester Wheatsheaf, 1992.

[37] J. C. Harsanyi, "Games with incomplete information played by "bayesian" players, i–iii part i. the basic model," *Management science*, vol. 14, no. 3, pp. 159–182, 1967.

[38] N. D. Duong, A. Madhukumar, and D. Niyato, "Stackelberg bayesian game for power allocation in two-tier networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2341–2354, 2016.

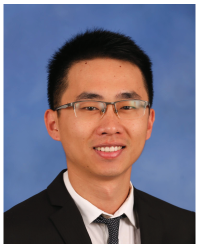
[39] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[40] J. Bethencourt, "Advanced crypto software collection," 2006. [Online]. Available: http://hms.isi.jhu.edu/acsc/libpaillier/

[41] F. P. Boscoe, T. O. Talbot, and M. Kulldorff, "Public domain small-area cancer incidence data for new york state, 2005-2009," *Geospatial health*, vol. 11, no. 1, p. 304, 2016.

[42] "Satscan," 2005. [Online]. Available: https://www.satscan.org/

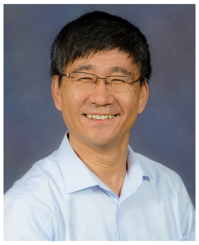
[43] J. Liu, C. Zhang, B. Lorenzo, and Y. Fang, "Dpavatar: A real-time location protection framework for incumbent users in cognitive radio networks," *IEEE Transactions on Mobile Computing*, vol. 19, no. 3, pp. 552–565, 2019.



Jianqing Liu (M'18) received the Ph.D. degree from The University of Florida in 2018 and the B.Eng. degree from University of Electronic Science and Technology of China in 2013. He is currently a tenure-track assistant professor at the Department of Electrical and Computer Engineering at University of Alabama in Huntsville. His research interest is wireless networking, mobile health, security and privacy. He received the U.S. National Science Foundation Career Award in 2021. He is also the recipient of several best paper awards including the 2018 Best Journal Paper Award from IEEE Technical Committee on Green Communications & Computing (TCGCC).

Chi Zhang (M'11) received the B.E. and M.E. degrees in electrical and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 1999 and 2002, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2011. He joined the School of Information Science and Technology, University of Science and Technology of China, as an Associate Professor in 2011. His research interests include the areas of network protocol design and performance analysis and network security.

Kaiping Xue (SM'15) received the bachelor's degree from the Department of Information Security, University of Science and Technology of China (USTC) in 2003, and the Ph.D. degree from the Department of Electronic Engineering and Information Science (EEIS), USTC in 2007. From May 2012 to May 2013, he was a Post-Doctoral Researcher with the Department of Electrical and Computer Engineering, University of Florida. He is currently a Professor with the School of Cyber Security and the Department of EEIS, USTC. His research interests include next-generation internet architecture design, transmission optimization, and network security. He is an IET fellow. He serves on the Editorial Board of several journals, including the IEEE Transactions on Dependable and Secure Computing (TDSC), the IEEE Transactions on Wireless Communications (TWC), and the IEEE Transactions on Network and Service Management (TNSM). He has also served as a Guest Editor for IEEE Journal on Selected Areas in Communications (JSAC), and a Lead Guest Editor for IEEE Communications Magazine and IEEE Network.



Yuguang Fang (F'08) received the M.S. degree from Qufu Normal University, Shandong, China, in 1987, the Ph.D. degree from Case Western Reserve University in 1994, and the Ph.D. degree from Boston University in 1997. He joined the Department of Electrical and Computer Engineering, University of Florida in 2000, and has been a Full Professor since 2005 and a Distinguished Professor since 2019. He holds the University of Florida Research Foundation (UFRF) Professorship from 2017 to 2020 and 2006 to 2009, the University of Florida Foundation Preeminence Term Professorship since 2019, and the University of Florida Term Professorship since 2017. He received the U.S. National Science Foundation Career Award in 2001, the Office of Naval Research Young Investigator Award in 2002, the 2019 IEEE Communications Society AHSN Technical Achievement Award, the 2015 IEEE Communications Society CISTC Technical Recognition Award, the 2014 IEEE Communications Society WTC Recognition Award, and the Best Paper Award from the IEEE International Conference on Network Protocols (ICNP) in 2006. He has also received the 2010–2011 UF Doctoral Dissertation Advisor/Mentoring Award, the 2011 Florida Blue Key/UF Homecoming Distinguished Faculty Award, and the 2009 UF College of Engineering Faculty Mentoring Award. He was the Editor-in-Chief of the IEEE Transactions on Vehicular Technology from 2013 to 2017, and the IEEE Wireless Communications from 2009 to 2012, and serves/served on several editorial boards of journals, including Proceedings of IEEE since 2018, ACM Computing Surveys since 2017, the IEEE Transactions on Mobile Computing from 2003 to 2008 and from 2011 to 2016, the IEEE Transactions on Communications from 2000 to 2011, and the IEEE Transactions on Wireless Communications from 2002 to 2009. He has been actively participating in conference organizations, such as serving as the Technical Program Co-Chair for the IEEE INFOCOM'2014 and the Technical Program Vice-Chair for the IEEE INFOCOM'2005. He is a fellow of the American Association for the Advancement of Science (AAAS).