2018 IEEE Symposium on Privacy-Aware Computing

A Cloud-based Secure and Privacy-Preserving Clustering Analysis of Infectious Disease

Jianqing Liu, Yaodan Hu, Hao Yue, Yanmin Gong, Yuguang Fang, *Fellow, IEEE*

***Abstract*—The early detection of where and when fatal infectious diseases outbreak is of critical importance to the public health. To effectively detect, analyze and then intervene the spread of diseases, people's health status along with their location information should be timely collected. However, the conventional practices are via surveys or field health workers, which are highly costly and pose serious privacy threats to participants. In this paper, we for the first time propose to exploit the ubiquitous cloud services to collect users' multi-dimensional data in a secure and privacy-preserving manner and to enable the analysis of infectious disease. Specifically, we target at the spatial clustering analysis using Kulldorf scan statistic and propose a key-oblivious inner product encryption (KOIPE) mechanism to ensure that the untrusted entity only obtains the statistic instead of individual's data. Furthermore, we design an anonymous and sybil-resilient approach to protect the data collection process from double registration attacks and meanwhile preserve participant's privacy against untrusted cloud servers. A rigorous and comprehensive security analysis is given to validate our design, and we also conduct extensive simulations based on real-life datasets to demonstrate the performance of our scheme in terms of communication and computing overhead.**

***Index Terms*—public health, clustering analysis, Kulldorf scan statistic, group signature, identity-based encryption, secure multi-party computation.**

I. Introduction

The analysis and surveillance of infectious disease is the cornerstone of the modern public health. It has become the international top priority to prevent the disease and optimize the health of the population [1]. The infectious disease could be rare (e.g., plague, Ebola) or recurrent (e.g., influenza), natural or intentional (e.g., bio-terrorism), but, regardless, it causes devastating consequences for the people and economies. In 2013, over 200,000 Canadians got infected by highly contagious diseases while 8,000 of them died as a result [2]. In October 2001, the anthrax attack [3], which was intentionally launched within the United States, killed 5 people and infected 17 others. Obviously, the impact of infectious diseases is immense and it has fueled demand for prospective systems for early detection and intervention of disease outbreaks and spread. Distinct from retrospective analysis, early (or real-time) identification of infectious disease attracts more research attentions, as it is expected to improve the survival rate of infected patients and to avoid more severe societal and economical loss due to disease dissemination [4].

Among all the practices for infectious disease surveillance, the spatial clustering analysis is viewed as the essential one, while the individual's health and location data have been deemed to be the two pillars for its application [5]. Specifically, by analyzing the health and location information of potential patients, epidemiologists could identify geographical disease clusters at the early stage of the infectious disease outbreak. The public resource (e.g., field health workers or antibiotic prophylaxis) could then be allocated to prevent its further dissemination. In recent years, the deployment of disease detection systems has become a reality. For instance, Brazil launched a mobile infectious disease surveillance project targeted for the Dengue fever - a constant threat to local residents [6]. The residents' data is collected on daily basis by field health workers who send individual's location information along with the respective survey results of their health status back to the server for early disease detection. MIT Lincoln Laboratory also deployed a similar project called Biological-Agent Correlation Tracker (BACTracker) [7] aiming to mitigate bio-terror attacks, which collects volunteers' location and syndromic records (e.g., respiratory, gastrointestinal or other fever-associated symptoms) periodically. A statistical analysis is then carried out after the data collection to pinpoint geographical disease clusters. Other systems like Gripenet started in 2005 in Portugal [8], FluTracking initiated by Australia in 2006 [9], Influweb started in 2007 in Italy [10], are widely adopted by different countries to collect volunteers' information for effective infectious disease analysis.

Ideally, to calculate the most timely and fine-grained infectious disease clusters, epidemiologists desire as much individual's information (i.e., health and location records) as possible, and at the highest possible level of precision. The aforementioned participatory-based systems, however, bear the weakness such as the poor data validation, limited representativeness, and unreliable participation rate [11]. Specifically, most of the above systems collect data weekly; the majority of participants are women [12], [13]; reporting fidelity is quite variable with some only reporting sporadically over time [12]; 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

J. Liu is with Department of Electrical and Computer Engineering at University of Alabama in Huntsville, Huntsville, AL, 35899 Email: jianqing.liu@uah.edu

Y. Hu and Y. Fang are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA e-mail: yaodan.cindy.hu@gmail.com, fang@ece.ufl.edu

H. Yue is with the Department of Computer Science, San Francisco State University, San Francisco, CA 94132 USA e-mail: haoyue@sfsu.edu

Y. Gong is with the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK 74078 USA e-mail: yanmin.gong@okstate.edu

This work was partially supported by National Science Foundation under grants IIS-1722791.

for a variety reasons, some related to disease severity, some attributed to socio-demographic differences [14], and others for privacy concerns (e.g., unwanted intrusive marketing, risk of legal or compliance exposure, etc. [15]). Therefore, a more timely, pervasive, secure and privacy-preserving data collection system is in dire need as an additional or supplemental data source for a more comprehensive clustering analysis of infectious disease.

In this paper, we are inspired to leverage the existing ubiquitous cloud services - in particular the location-based services by cloud servers (e.g., Google, Yelp, etc.) and health-related services by cloud servers (Intel's Health Guide, GE's QuiteCare, etc.) - to collect the timely, ubiquitous and representative statistic data so as to facilitate the clustering analysis of infectious diseases. In the meantime, we attempt to guarantee the normal cloud services to users, to preserve user's privacy and to ensure the system security. To assure the applicability of this proposal, a viable practice is to send users consent forms notifying them that statistical rather than the fine-grained records will be utilized for scientific study (i.e., public health), whereas a certain monetary incentives could be provided as well. The incentive design is not within the scope of this paper but we will investigate it in the future research.

The major contributions we have made in this paper are summarized as follows.

- To the best of our knowledge, this is the first work proposing to exploit the ubiquitous multi-cloud platforms for the spatial clustering analysis of infectious diseases. This proposal will significantly enhance/complement the conventional data collection and disease analysis paradigm.
- We develop a secure data collection protocol based on anonymous group signature to protect system security against double registration attacks while preserving participants' location and health privacy.
- We design a secure multi-party computation scheme to ensure the untrusted entities can only get statistic (i.e., sum value) but are oblivious of each individual's data.

The remainder of the paper is organized as follows: Section II describes the system model, security assumptions and design objectives. Section III introduces the preliminaries for our later design. Section IV presents our scheme design which is break down into different phases. The security proof is given in Section V and the performance evaluation of our scheme is then showed in Section VI. Finally, Section VII concludes the paper.

II. System Model and Design Objectives

In this section, we give a high-level discussion of the system model for spatial epidemiology analysis, the threat and security model, and our design objectives.

A. System Model

There are five entities in our system: trusted authority (TA), public health office (PHO), location-based service cloud server (LC), health service cloud server (HC) and users, as shown in Fig.1. Users are capable of interacting with the LC and HC through off-the-shelf technologies such as their GPS-enabled mobile phones, wearable devices (e.g., Apple watch), respectively. In this respect, the LC and HC can correspondingly collect user's location and health information, which are later exploited by the PHO for the spatial epidemiology analysis. More explanations of each entity in the system are given below.

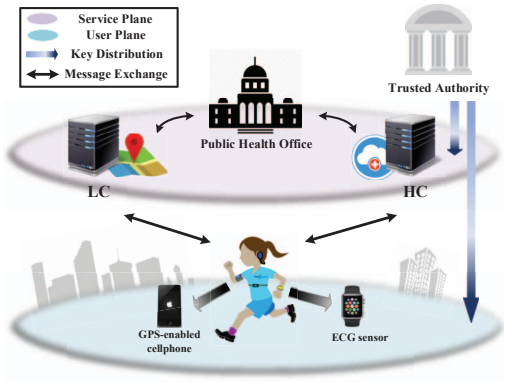

Figure 1: System model of the infectious disease analysis exploiting two clouds.

- **Trusted Authority (TA)** is an entity that bootstraps the system, creates and distributes credentials for legal entities in the system. TA is also responsible for solving disputes, revoking compromised entities when misbehavior are reported.
- **Public health office (PHO)** is a government entity in a geographic region (e.g., city level or state level) who conducts analysis and control over the spread of epidemic diseases. To do so, PHO collects population's mobility (i.e., location) and health information periodically (e.g., in a daily basis) via queries to the LC and HC, respectively.
- **Location-based service cloud server (LC)** is a company-operated cloud server such as Google or Yelp that collects users' location information to provision rich location-based services (LBS). On top of the existing services, our system allows PHO to access LC to perform epidemiology analysis, but in a privacy-preserving manner.
- **Health service cloud server (HC)** is similar to LC in the sense that it is an enterprise cloud server that has powerful computation and storage capabilities. HC collects users' rich set of health information such as respiratory, gastrointestinal symptoms, etc. to analyze users' health condition using sophisticated machine learning models. In addition to that, HC is also made available to PHO under extreme circumstances like the outbreak of epidemic diseases, but each user's health information should be kept private.
- **User** first registers to the TA at the system initialization phase to obtain valid credentials. Users then interact with the LC and HC using their mobile phones and wearable devices to obtain LBS and health services, respectively.

B. Security Model

TA is fully trusted by all other entities in the system and is assumed not compromised. LC and HC are honest-but-curious, i.e., they honestly follow the protocol but are curious about users' location and health information, respectively. Moreover, LC and HC may be operated by one enterprise (e.g., Google) so their records could be combined to predict a user's future health condition through his/her social contacts with other infected users, which may result in unwanted intrusive marketing or denial of insurance coverage for that particular user. PHO is assumed honest-but-curious in the sense that it honestly conducts statistic analysis of the spread of epidemic diseases but are curious of individual's location and health information for purposes like segregating infected patients, which however might be against users' willingness and compromise their privacy. Last but not least, the users in this system are not trusted. They may launch sybil attacks to mislead PHO's statistic analysis to either cause panic in an uninfected region or reduce PHO's awareness of an infected area, for the purpose of bio-terrorism or gaining commercial advantages.

C. Design Objectives

Based on the discussion of the prior security model, we present the design objectives as follows.

- *Individual data privacy*. The data privacy indicates the confidentiality of the location and health data. Each user's data privacy should be protected against potential adversaries and the aforementioned curious entities.
- *Usability*. The submitted data should allow PHO to conduct statistical analysis of epidemic diseases and LC and HC to provision corresponding services to users.
- *Data Verifiability and user accountability*. LC and HC should be able to authenticate users and their submitted data in order to avoid them misleading PHO through injecting falsified data. The system is also expected to revoke misbehaved users and reject bogus data.
- *Efficiency*. The spatial epidemiology analysis deals with a significant amount of user records (e.g., hundreds of thousands) in a city or state level. The protocol should be efficient enough to provide timely analysis and response for the epidemic disease control.

III. Preliminaries

A. Kulldorff Spatial Scan Statistic [16]

The Kulldorff spatial scan statistic, firstly proposed in [16] 1997, now becomes one of the most powerful tools in performing clustering analysis to detect small clusters in a large location dataset. It finds a great potential in the application of spatial epidemiology surveillance to discover the small spatial regions (e.g., a school or a shopping mall) of significantly elevated disease density [17]. The core idea of the Kulldorff spatial scan statistic is to detect the overdensity region than any other statistics. The following describes how the Kulldorff spatial scan statistic works.

An surveillance region G is firstly divided into subareas (e.g., district or county) $\{s_1, s_2, ..., s_K\}$ of any arbitrary level of fine-grain, such that $G = \bigcup_{i=1}^{K} s_i$. For demonstrative purposes, Fig.2 shows the geographic map of New York State, which is sliced into 62 subareas according to the county divisions. PHO then collects the statistics of the total disease case count and population in each subarea, denoted as $\{c_1, c_2, ..., c_K\}$ and $\{p_1, p_2, ..., p_K\}$, respectively. In so doing, PHO keeps a record of 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 G. The Kulldorff spatial scan statistic is then applied to search all the cluster of neighbouring subareas to find abnormal ones with disease overdensity. In specific, suppose $\{S_1, S_2, ..., S_M\}$ is the set of all the possible cluster of adjacent subareas, each of which has disease case count C_j and population P_j. As an example, Fig.2 shows a cluster area $S_1 = s_{36} \cup s_{37} \cup s_{38}$. The Kulldorff spatial scan statistic then proceeds to calculate the respective cluster density D_j as

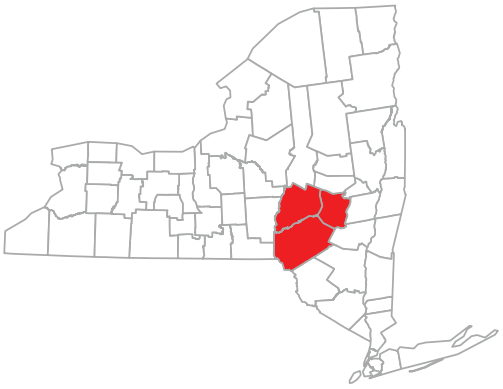

Figure 2: A demonstrative example for Kulldorff scan statistic: 62 counties in New York State while three of them forming a spatial cluster as the result of running Kulldorff scan statistic.

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

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

In so doing, PHO can detect the maximum density $mrd = \max_{S_j \in G} D_j$ and the corresponding cluster $mdr = \arg\max_{S_j \in G} D_j$ in the region G. To evaluate whether this cluster is statistically significant or spatial overdensity in terms of the disease case count, the Kulldorff spatial scan statistic assumes c_i following inhomogeneous Poisson processes and a randomization testing approach is conducted to test whether mdr is statistically significant. To be specific, PHO firstly generates R of random replications of the region G. Each replica has the same underlying populations $\{p_1, p_2, ..., p_K\}$ as the benchmark G, but assumes a uniform disease rate $q_{rp} = \frac{C_{tot}}{P_{tot}}$ for all the subarea $\{s_1, s_2, ..., s_K\}$. Then, for each replica G', PHO draws c_i randomly from an inhomogeneous Poisson distribution with mean $q_{rp} p_i$, and finds the mdr of G'. The statistical significance (i.e., p-value) is then calculated as the number of replications having $mrd(G') \geq mrd(G)$ divided by the total number of replications R. Normally, the cluster is

considered as the outlier or being statistically significant when $p \leq 0.05$.

B. Bilinear Pairing

Bilinear pairing-based cryptography has attracted great interests from the security community as it enables several innovative designs such as IBC [18]. Although many groups with a useful bilinear map area based on elliptic curve, our definitions are abstract and we follow the notion of Boneh *et al.* [18]. Let $\mathbb{G}_1$ and $\mathbb{G}_2$ being two multiplicative cyclic groups of prime order p, and g_1 and g_2 be the generator of $\mathbb{G}_1$ and $\mathbb{G}_2$, respectively, and ψ be an efficiently computable bilinear map from $\mathbb{G}_2$ to $\mathbb{G}_1$ such that $\psi(g_2) = g_1$. The following property holds true:

- **Bilinearity**: $e\left(P^a, Q^b\right) = e(P,Q)^{ab}$ for all $P \in \mathbb{G}_1$, $Q \in \mathbb{G}_2$ and $a, b \in \mathbb{Z}_p^*$.
- **Non-degeneracy**: $e\left(g_1, g_2\right) \neq 1$.
- **Computability**: $e\left(P, Q\right)$ can be computed efficiently for any $P \in \mathbb{G}_1$, $Q \in \mathbb{G}_2$.

In this paper, we consider the bilinear map $e : \mathbb{G}_1 \times \mathbb{G}_2 \rightarrow \mathbb{G}_T$, where $\mathbb{G}_1 \neq \mathbb{G}_2$, although one could set them equal.

IV. The Secure and Privacy-Preserving Protocol for Infectious Disease Analysis

In this section, we present the details of our protocol design by exploiting two clouds (i.e., LC and HC) to ensure PHO can conduct the spatial clustering analysis of infectious diseases in a secure and privacy-preserving manner.

A. Protocol Overview

Compared with the existing disease surveillance projects such as the BACTrack system relying on participatory volunteers [7] or the system supported by the field workers or the family physicians [15], our system depends on the running cloud services which are ubiquitous in everyone's cyber-life. The major drawbacks of prior disease surveillance systems include the limited number of participants, the huge system operating costs, and the inefficient, insure and non-private data collection process; whereas our system can achieve obvious advantages in these respects. However, there are several unique design challenges in our system as well which was described in Section II. We thereby first give a high-level overview of our secure and privacy-preserving protocol.

At the moment of infectious disease outbreak, the government or PHO can either initiate a request to the LC and HC to collect users' location and health information within a response time (i.e., 10:00am to 10:15am of Jan.8th 2018 in NYC) or directly query the existing records in the LC and HC. In this work, we apply the former model to guarantee the freshness of the collected data. Then, within the data collection phase, mutual authentication should be carried out between users and LC & HC so that the system security (i.e., preventing double registration, revoking misbehaved users) could be ensured. In the meantime, users' privacy must be preserved as well. After the data collection phase, PHO is only limited to query LC and HC for the statistic data (i.e., sum) to conduct the Kulldorf scan statistic analysis.

In what follows, we will give the in-depth discussions of our protocol.

B. Privacy-Preserving Authentication With Resilience to Sybil Attacks

In this design phase, the objective is two-fold: LC and HC need to authenticate users to ensure that they are authorized users (i.e., not on the revocation list) and each of them only submits a single data tuple (i.e., location and health data) within the response time session; whereas users demand the authentication process not to reveal their real identities while LC & HC not being able to link their multi-dimensional data through collusion. The proposed solution is inspired by an efficient group signature scheme [19], which serves as the basis to achieve the aforementioned objectives. Concretely, our scheme consists of three parts: (1) key generation, (2) signature, (3) verification.

(1) **KeyGen**: Given the security parameter κ, TA firstly bootstraps the system by generating the tuple $(p, g_1, g_2, \mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_T, e)$. TA also chooses two secure cryptographic hash functions H_0 and H, with respective ranges $\mathbb{G}_2^2$ and $\mathbb{Z}_p^*$. Next, TA randomly picks $\gamma \in \mathbb{Z}_p^*$ and sets $w = g_2^{\gamma}$. Using γ, for each user i, TA selects a random number $x_i \in \mathbb{Z}_p^*$ such that $x_i + \gamma \neq 0$, and then sets $A_i = g_1^{1/(x_i+\gamma)}$. In so doing, TA could publish the public parameter $pub\_para = (p, g_1, g_2, \mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_T, e, \omega, H, H_0)$, where the partial tuple (g_1, g_2, ω) in [19] is denoted as the *group public key (gpk)*. For each user i, TA also distributes its *group secret key (gsk)*, $gsk = (x_i, A_i)$, through a secure channel.

(2) **Sign**: Suppose cloud servers have their unique identifiers as $lcid$ and $hcid$, respectively, and every response time session is assigned with an identifier sid as well, $lcid$, $hcid$ and sid are made public and users will then conduct the following steps to anonymously authenticate themselves. Note that for a clear presentation, we only exhibit the authentication process between users and LC while the authentication between users and the HC can be conducted in a similar manner. However, we shall demonstrate in the later session that LC and HC cannot link the user through this process.

1) Generate a tuple in $\mathbb{G}_2$ from H_0, and then compute their images in $\mathbb{G}_1$ from ψ:
$$(\widehat{u}, \widehat{v}) \leftarrow H_0(gpk, sid, lcid) \in \mathbb{G}_2,$$
$$u \leftarrow \psi(\widehat{u}) \in \mathbb{G}_1,\ v \leftarrow \psi(\widehat{v}) \in \mathbb{G}_1.$$
2) Select a random number $\alpha \in \mathbb{Z}_p^*$, and then compute:
$$T_1 = u^{\alpha},\ T_2 = A_i v^{\alpha}.$$
3) Set $\delta = \alpha x_i$ and choose random numbers $r_{\alpha}, r_x, r_{\delta} \in \mathbb{Z}_p^*$, then compute:
$$\begin{cases} R_1 = u^{r_{\alpha}} \\ R_2 = e(T_2, g_2)^{r_x} \cdot e(v, \omega)^{-r_{\alpha}} \cdot e(v, g_2)^{-r_{\delta}} \\ R_3 = T_1^{r_x} \cdot u^{-r_{\delta}} \end{cases}$$
4) Compute a challenge value from H:
$$c \leftarrow H(gpk, sid, T_1, T_2, R_1, R_2, R_3) \in \mathbb{Z}_p^*.$$

5) Compute $s_\alpha = r_\alpha + c\alpha$, $s_x = r_x + cx_i$ and $s_\delta = r_\delta + c\delta$, and construct the following authentication message:

$$\sigma \leftarrow (sid, T_1, T_2, c, s_\alpha, s_x, s_\delta).$$

(3) **Verify**: Upon receiving users' authentication message σ, the cloud server proceeds in three phases: it first examines the validity of the authentication message σ; it then checks if this user is on the revocation list (RL); finally it ensures this message is not from the double registration or sybil users. The cloud servers only accept users' data (i.e., location and health data) if these conditions hold.

1) Compute the following tuple:

$$\begin{cases} \overline{R_1} = u^{s_\alpha} \Big/ T_1^c \\ \overline{R_2} = e(T_2, g_2)^{s_x} \cdot e(v, \omega)^{-s_\alpha} \cdot e(v, g_2)^{-s_\delta} \cdot \left[\dfrac{e(T_2, \omega)}{e(g_1, g_2)}\right]^c \\ \overline{R_3} = T_1^{s_x} \cdot u^{-s_\delta} \end{cases}$$

2) Check if the following holds:

$$c \stackrel{?}{=} H(gpk, sid, T_1, T_2, \overline{R_1}, \overline{R_2}, \overline{R_3}). \tag{2}$$

3) If so, for every $A \in RL$, check if the following holds to see if the user is on the revocation list:

$$e(T_2/A, \widehat{u}) \stackrel{?}{=} e(T_1, \widehat{v}). \tag{3}$$

4) If not, check if the following holds to avoid sybil attacks:

$$e(T_2, \widehat{u})/e(T_1, \widehat{v}) \stackrel{?}{=} e(T_2^{'}, \widehat{u}) \Big/ e(T_1^{'}, \widehat{v}). \tag{4}$$

In [19], it is proven that revocation check can be conducted locally (i.e., in LC and HC in this paper) via examining whether A is encoded in (T_1, T_2) through (3). Similarly, inspired by (3), we can leverage the relation that $e(A_i, \widehat{u}) = e(T_2, \widehat{u})/e(T_1, \widehat{v}) = e(T_2^{'}, \widehat{u}) \Big/ e(T_1^{'}, \widehat{v})$ always holds for the same user i possessing A_i within a response time session sid to check if the same user (or the sybil node) attempts to double register in the cloud servers. If misbehaved users are detected in this process, (T_1, T_2) shall be reported to the TA which will later update a new revocation list to the cloud servers.

Briefly, the core idea of our design is to allow the user to anonymously prove to cloud servers for possession of a *Strong Diffie-Hellman (SDH)* tuple, i.e., (x_i, A_i), via solving a challenge problem. And the check for double registration attack is also in line with this same intuition. For interested users, the security proofs (e.g., correctness) for this protocol can be found in [19], but we shall discuss in the later section that this design will also not allow the LC and HC to link a specific user if they collude.

C. Privacy-Preserving Data Collection

After the user authenticates itself to the cloud servers, its location and health data will be sent to the LC and HC, respectively. To preserve the data privacy, pseudonyms or data encryption/obfuscation are commonly adopted approaches. Depending on different application scenarios, for instance, some literature may consider cloud servers are limited to access the ciphertext [20], [21]. In this work, we leverage the pseudonym crypto-system and the cloud servers have users' plaintext data so that on one hand they could provision users services while on the other hand allow PHO to conduct infectious disease analysis.

The challenge of data collection phase comes from that the design should allow PHO to link users' multi-dimensional (i.e., location and health) data while achieving unlinkability between LC and HC. Inspired by ID-based encryption (IBE) proposed by Boneh and Franklin in [18], we develop an IBE-based access control scheme to fulfill the objective. Suppose user i sends HC following message tuple (uid_i, h_i) where uid_i is its pseudonym and h_i is the health data. Then, the user encrypts this pseudonym using IBE scheme as $uid_i^{'} \leftarrow Enc_{IBE}(pp, \overline{pid}, uid_i)$ where pp is the public key generated by TA and $\overline{pid}$ represents the cryptographic hash of PHO's identity pid. After that, the user sends LC following message tuple $(uid_i^{'}, loc_i)$ where loc_i is its location information. In so doing, LC and HC cannot link any two data records in their database due to the probabilistic encryption nature of IBE; whereas PHO could associate them by decrypting the identifier from LC $uid_i \leftarrow Dec_{IBE}(\overline{pid}, uid_i^{'})$ using the private key obtained from the TA.

D. Privacy-Preserving Data Query

By the end of the response time session, LC and HC have collected users' location and health information, respectively. To enable the infectious analysis using Kulldorf scan statistic, PHO needs to query the LC and HC to obtain the population P_j and the count of infected users C_j in each geographical grid j. Our objective is to let the PHO only access the sum value (i.e., P_j and C_j) without revealing individual user's location and health information. Besides, the query process should be as computationally efficient as possible since we may deal with tens of thousands of users' data records within a city. This naturally leads us to design a privacy-preserving batch query scheme, but we have to ensure that the batch query would not leak additional information to LC or HC to link any users.

In this design phase, the PHO starts with sending queries to LC for the list of users' pseudonyms $uid_i^{'}$ in every geographical grid. The PHO only knows a batch of users in a rough geographical area but is oblivious of where specifically a user is. Then, the PHO solicits the IBE private key from the TA by proving its identity credentials in order to decrypt the users' pseudonyms which was previously encrypted as in Section IV-C. Now, the PHO has the pseudonym list $\boldsymbol{uid}$ that is exactly matched with that is stored in HC. On the other hand, HC may employ some models, like the Naive Bayes in [21], to evaluate users' health status based on their submitted health records h_i. Hence, we suppose the HC now maintains a health vector of 0s and 1s representing whether a user is infected (i.e., 1) or not (i.e., 0); whereas the indices of this vector are users' respective pseudonyms uid_i. To privately and effectively retrieve the count of infected users in every geographical grid, the PHO and HC need to conduct the following computation task. For demonstrative purposes, we show a toy example in Fig.3. Suppose the PHO constructs a query matrix $\boldsymbol{Q}$ containing

users' residence in each geographical grid. An integer value 1 represents the user is within that grid; 0 otherwise. We also assume users are not located in the boundary of grids so every user's residence can be uniquely captured by a specific geographical grid. On the other hand, HC keeps users' infected status in a data vector $\boldsymbol{H}$. Note that the user list in $\boldsymbol{Q}$ and $\boldsymbol{H}$ could be mismatched if some users did not submit the pair of their location and health information in the data collection phase. Hence, for accurate analysis, our design is expected to filter the mismatched entries in a privacy-preserving manner. In the end, the PHO could derive the case count vector $\boldsymbol{CNT}$ in each geographical grid via the inner product of $\boldsymbol{Q}$ and $\boldsymbol{H}$.

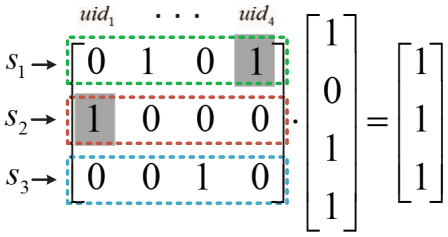

Figure 3: A toy example for the batch query: PHO's query matrix contains 3 grids and 4 users; HC's database holds 4 users' infected status; and the inner product gives the count of infected users in each grid.

To preserve users' location privacy against the HC and the health privacy against the PHO, PHO's query matrix $\boldsymbol{Q}$ should be kept private to the HC and oppositely PHO should be oblivious to HC's health vector $\boldsymbol{H}$ as well. Therefore, our scheme is boiled down into a secure multiparty computation (SMC) [22], [23] design where the PHO and HC should collaboratively compute the inner product of $\boldsymbol{Q}$ and $\boldsymbol{H}$ in a privacy-preserving way.

Our design is inspired by the secure k nearest neighbour (kNN) scheme [24], which is to securely search the k nearest database records in the Euclidean distance between a data record $\boldsymbol{p}$ and a query vector $\boldsymbol{q}$. To adapt the Euclidean distance, every data record $\boldsymbol{p_i}$ and query record $\boldsymbol{q}$ is firstly extended to $(d+1)$-dimension where the $(d+1)$th element is set as $-0.5\|p_i\|^2$ and 1, respectively. Then, a random invertible matrix $\boldsymbol{M}$ of dimension $(d+1)\times(d+1)$ is used to encrypt the data and query record through the following equation

$$\begin{cases} \boldsymbol{p_i'} = (\boldsymbol{p_i}, -0.5\|\boldsymbol{p_i}\|^2)\cdot\boldsymbol{M} \\ \boldsymbol{q'} = \boldsymbol{M^{-1}}\cdot(\boldsymbol{q}, 1)^T \end{cases}$$

where it is proven in [24] that the scheme can determine whether $\boldsymbol{p_i}$ is closer to $\boldsymbol{q}$ than $\boldsymbol{p_j}$ is by comparing $(\boldsymbol{p_i'}-\boldsymbol{p_j'})\cdot\boldsymbol{q'}$ with 0. The reason is that to examine if $\sqrt{\|\boldsymbol{p_i}\|^2-2\boldsymbol{p_i}\cdot\boldsymbol{q}+\|\boldsymbol{q}\|^2} \geq \sqrt{\left\|\boldsymbol{p_j}\right\|^2-2\boldsymbol{p_j}\cdot\boldsymbol{q}+\|\boldsymbol{q}\|^2}$, it is equivalent to evaluate whether $\|\boldsymbol{p_i}\|^2-\left\|\boldsymbol{p_j}\right\|^2-2(\boldsymbol{p_i}-\boldsymbol{p_j})\boldsymbol{q} \geq 0$ or not. Using the secure kNN scheme in [24], we have the inner product of encrypted data record $\boldsymbol{p_i'}$ and query record $\boldsymbol{q_i'}$ equals $\boldsymbol{p_i}\cdot\boldsymbol{q}-0.5\|\boldsymbol{p_i}\|^2$ which can be further used to compare $\|\boldsymbol{p_i}\|^2-\left\|\boldsymbol{p_j}\right\|^2-2(\boldsymbol{p_i}-\boldsymbol{p_j})\boldsymbol{q}$ with 0 via calculating $(\boldsymbol{p_i'}-\boldsymbol{p_j'})\cdot\boldsymbol{q'}$. Furthermore, the work [24] also presents to apply random asymmetric splitting and add artificial dimensions to enhance the security of the secure kNN scheme.

However, we need to do some modifications on this conventional scheme as the random matrix $\boldsymbol{M}$ in our framework should be kept private to the query side (i.e., HC) and the PHO needs to derive the exact value of the inner product of $\boldsymbol{Q}$ and $\boldsymbol{H}$ instead of a relative number. Therefore, we propose a Key-Oblivious Inner Product Encryption (**KOIPE**) scheme, which bears a similar idea with [24], [25]. Our proposed scheme consists of the following components: (1) matrix synchronization, (2) query randomization, (3) data encryption, (4) key embedment, (5) inner product. To give a clearer picture, we also show the overall information exchange of the data query phase in Fig.4.

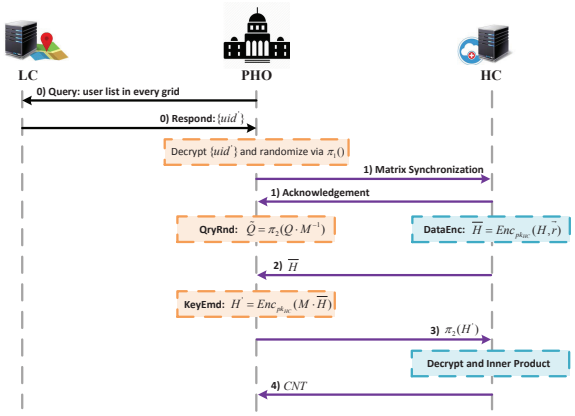

Figure 4: Three party information exchange diagram for the privacy-preserving data query.

(1)**MtxSync**: Before executing the secure inner product, the PHO and HC should "synchronize" or "align" their matrices so that any mismatched records could be eliminated. Instead of sorting user list $\boldsymbol{uid}$ according to their resided location grids, PHO applies a permutation mechanism $\boldsymbol{\pi_1}$ to randomize the sequence of the user list, which is then sent to HC to filter mismatched records and to re-order the health record $\boldsymbol{H}$.

(2)**QryRnd**: PHO selects a random invertible matrix $\boldsymbol{M}$ of size $N\times N$, to encrypt the query matrix into the following form $\overline{\boldsymbol{Q}} = \boldsymbol{Q}\cdot\boldsymbol{M^{-1}}$. To enhance the security level, we further randomize the encrypted query matrix using another permutation mechanism $\boldsymbol{\pi_2}$ and then send the transformed encrypted query matrix $\widetilde{\boldsymbol{Q}}$ (i.e., $\widetilde{\boldsymbol{Q}} = \boldsymbol{\pi_2}(\overline{\boldsymbol{Q}})$) to the HC. Note that applying another round of permutation is to further randomize the matrix while we are aware of the fact that the same permutation on two vectors does not change the inner product of them.

(3)**DataEnc**: We resort to the Paillier cryptosystem [26] in our scheme to support the additively homomorphic encryption, so the HC firstly obtains the key pair as (pk_{HC}, sk_{HC}) from the TA. Then, HC encrypts its health data vector using the public key pk_{HC} into $\overline{\boldsymbol{H}}$, which is then sent back to PHO. Specifically, for each record h_i in $\boldsymbol{H}$, the encryption operates

as $\overline{h_i} = Enc_{pk_{HC}}(h_i, r_i)$ where $Enc_{pk_{HC}}()$ is the encryption function in the Paillier cryptosystem and r_i is a random number selected in correspondence with the ith record of $\boldsymbol{H}$. Clearly, it is in the nature of probabilistic encryption so every record of $\boldsymbol{H}$ will not be encrypted into the same value. Interested reader can refer to [26] for more detailed descriptions.

(4)**KeyEmd**: Our intention here is to embed (or encrypt) the random invertible matrix $\boldsymbol{M}$ in $\overline{\boldsymbol{H}}$. We attempt to leverage the property of the Paillier cryptosystem in providing the additive homomorphism, which is described as 1) $Enc_{pk}(m_1, r_1) \times Enc_{pk}(m_2, r_2) = Enc_{pk}(m_1 + m_2, r_1 r_2)$ and 2) $Enc_{pk}(m_1, r_1)^k = Enc_{pk}(k \times m_1, r_1^k)$, to embed $\boldsymbol{M}$. Specifically, PHO need to calculate $\boldsymbol{H}' = Enc_{pk_{HC}}(\boldsymbol{M} \cdot \overline{\boldsymbol{H}})$ and to do so, the following arithmetic computation for each element in $\overline{\boldsymbol{H}}$ should be executed:

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

Then, PHO applies the same permutation $\boldsymbol{\pi_2}$ as presented in Step 2 to randomize the $\boldsymbol{H}'$, which is then sent back to HC along with the encrypted query matrix $\widetilde{\boldsymbol{Q}}$.

(4)**InPrd**: After receiving the two matrices from PHO in Step 4, HC first decrypts the health data vector using the secret key sk_{HC} into $\widetilde{\boldsymbol{H}}$ and then computes the inner product to derive the count of infected users as $\boldsymbol{CNT} = \widetilde{\boldsymbol{Q}} \cdot \widetilde{\boldsymbol{H}^T}$. Then, HC sends the vector $\boldsymbol{CNT}$ back to the PHO, which concludes the whole process.

Upon obtaining the disease case count c_j and the respective population (i.e., number of participants) p_j for every geographical grid area s_j, PHO first calculates the cluster density D_j according to (1) and then runs the randomization test to search for the spatial clusters that exhibit the statistical significance. These discovered clusters will be marked as infectious areas requiring follow-up actions from the PHO.

V. Security Analysis

Theorem 1. *Our scheme preserves users' privacy against the LC, HC and PHO.*

Proof. In the data collection phase, we require users to authenticate themselves to LC and HC to avoid double registration attacks. The authentication is a process of knowledge proof which means by rewinding a prover it is possible to extract an SDH pair (i.e., (x, A)) but the verifier (i.e., HC and LC) is oblivious of the identity of the user. The only exception is that for the revoked user, the verifier is aware of whom he is interacting with but this information will not help the verifier deduce additional knowledge about other legitimate users as (x, A) is generated by the TA separately.

Besides, since A is ElGamal-encrypted in (T_1, T_2) and $(\widehat{u}, \widehat{v})$ is publicly known, one nature concern is whether the LC and HC can collude by sharing (T_1, T_2) to link a specific user through the authentication process. Firstly, the LC and HC cannot find the arithmetic relation between $(T_{1,HC}, T_{2,HC})$ and $(T_{1,LC}, T_{2,LC})$ due to the blinding number α for which the LC/HC has to solve a Discrete Logarithm (DL) problem. Secondly, $e(T_{2,HC}, \widehat{u_{HC}})/e(T_{1,HC}, \widehat{v_{HC}}) \neq e(T_{2,LC}, \widehat{u_{LC}})/e(T_{1,LC}, \widehat{v_{LC}})$ as $e(A, \widehat{u_{HC}}) \neq e(A, \widehat{u_{LC}})$, and even though $(\widehat{u}, \widehat{v})$ is known, A cannot be derived due to the hardness of solving a Computational Diffie-Hellman (CDH) problem. Moreover, when users submit their data, HC/LC cannot link users' pseudonyms as they need to have PHO's secret key in the IBE crypto-system.

Last but not least, in the data query phase, the PHO on one hand has no idea where a specific user is due to the batch query to the LC; on the other hand it cannot deduce each user's health status because HC employs the additively homomorphic encryption which is a probabilistic encryption approach. The HC, however, is oblivious of where each user is during the interaction with the PHO. The reason is that by observing PHO's $\widetilde{\boldsymbol{Q}}$ the 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. However, there lacks sufficient information to solve the M thus the HC cannot obtain users' relative location with each other through interacting with PHO. After getting the sum of disease count back from HC, we argue that each user's health information is "hidden in the crowd" (i.e., sum statistic) unless there is only one user in a geographical grid area.

Note that our privacy-preserving design is relied on the cryptographic pseudonym system. Although some research findings show that the pseudonym-based approach could fail for individual data privacy in some contexts if the adversary has side information to re-identify or de-anonymize the user. We do not see this is the case in our scenario but if this truly happens we could easily employ certain randomization approaches (e.g., differential privacy) to perturbate the location and health data. ∎

Theorem 2. *Our scheme is resilient to the double registration attack.*

Proof. To let HC and LC accept the data, a malicious user has to pass the authentication process by submitting the (T_1, T_2) tuple in the signature message. For a given time session, $(\widehat{u}, \widehat{v})$ is fixed and known to HC and LC. To bypass the double-registration verification in (4), the malicious user needs to construct another tuple (T_1', T_2') such that $e(T_2, \widehat{u})/e(T_1, \widehat{v}) \neq e(T_2', \widehat{u})\big/e(T_1', \widehat{v})$ while allowing LC and HC to solve the challenge in (2). To do so, the malicious user needs to compose another group secret key (x', A'), for which it has to derive the secret seed γ from the parameter in the group public key ω but it is a DL problem. Therefore, a malicious user cannot pass the authentication process and meanwhile double register its record in servers. ∎

VI. Performance Evaluation

In this section, we attempt to evaluate the incurred storage and computing overhead of our scheme for each entity, namely the user, LC, HC and PHO. Firstly of all, we put forward details of the simulation setup.

A. Simulation Setup

We use a workstation of 3.2GHz Intel(R) Core(TM) i3 CPU and 8GB memory to emulate the LC's, HC's and PHO's computing facility. We implement our security mechanism using the JPBC library [27] with Type D159 pairing internal to realize short group signature in our design. Specifically, it offers 160-bit prime order p, 159-bit length for elements in group $\mathbb{G}_1$ and is equivalent to 954 bits Discrete Logarithm security. Besides, we employ 2048-bit modulus as the secret key length, as with the RSA, in the Paillier cryptosystem.

Furthermore, we exploit two real-life datasets, namely the lung cancer incidence in New York State [28] and the birth defect data in New York State [29]. The former dataset is constructed by collecting 67,217 tumor incidences from 2005-2009 out of an average of 19.34 million population covering 13,848 spatial groups. The latter dataset contains 1,237,189 new born children from 2005-2009 and 24,940 of them have the congenital malformations. This dataset is geographically organized by the ZIP code (1,600 and 1,143 after aggregation).

B. Communication Overhead Analysis

We first evaluate the communication overhead on the user side. In the authentication phase, the user generates a signature message σ containing two elements of $\mathbb{G}_1$, four elements of $\mathbb{Z}_p^*$, and one short session indicator. Given the prior simulation setup, the signature length of σ is equal to $l_{1,u} = 159 \times 2 + 160 \times 4 + 16 = 974$ bits where we assume the length of session ID is 16bits but it depends on how granular the PHO attempts to collect the data. In the data submission phase, the user generates a pseudonym and then encrypts it using PHO's ID, which altogether counts for two elements of $\mathbb{Z}_p^*$. Thus, the length of the two pseudonyms is $l_{2,u} = 160 \times 2 = 320$ bits. Here we neglect the data payload (i.e., location and health data) as they are inevitable regardless of what security mechanisms are developed, so the incurred extra data size for communications between one user and LC/HC is $l_u = l_{1,u} + l_{2,u} = 974 + 320 = 1,294$ bits. Note that we do not measure the latency as the communication channel conditions (e.g., Wi-Fi or cellular) are difficult to obtain. Therefore, we use the metric of transmitted data size to indicate the communication overhead.

For the communications on the server side, the overhead for PHO's location query to LC is too negligible to be counted. The response message from LC to PHO, however, has the data size of $l_{lc} = N \times 160$ bits where N is the number participant users and each user's pseudonym length is 160 bits. The communications between HC and PHO, on the other hand, is more complicated. Specifically, the communication overhead of PHO in the phase of **MtxSync** is $l_{1,PHO} = N \times 160$ bits whereas HC has roughly the same communication overhead. The PHO generates a randomized query matrix $\overline{\overline{\boldsymbol{Q}}}$ in **QryRnd** which is sent to HC. Suppose the element in $\overline{\overline{\boldsymbol{Q}}}$ is stored in double type of 64 bits; then the communication overhead for PHO is $l_{2,PHO} = N \times K \times 64$ bits. In the phase of **DataEnc**, every element in health vector $\boldsymbol{H}$ is encrypted via the additive homomorphic encryption scheme so the data size of each record in $\overline{\boldsymbol{H}}$ is 4096 bits which amounts to $l_{2,HC} = N \times 4096$ bits. The PHO then encodes the random invertible matrix $\boldsymbol{M}$ into $\overline{\boldsymbol{H}}$ in the **KeyEmd** phase and the data size of $\boldsymbol{H}'$ equals to $l_{3,PHO} = N \times 4096$ bits. After computing the inner product in the **InPrd** phase, the HC sends $\boldsymbol{CNT}$ back to PHO which contributes the communication overhead of $l_{3,HC} = K \times 64$ bits.

Notice that for communications between servers, there exists high speed channels (e.g., optic fibers) between them so the actual communication delay could be neglected. To give a clearer picture, we summarize the total communication overhead, measured in bits, for each entity at each process in Table I.

C. Computation Overhead Analysis

We first theoretically examine the computation complexity in performing our proposed security scheme. Denote T_{mul}, $T_{\exp}$, T_{par}, $\widehat{T_{mul}}$ and $\widehat{T_{exp}}$ as the time to compute one multiplication over $\mathbb{G}_1$, one exponentiation over $\mathbb{G}_1$, one pairing over $\mathbb{G}_T$, one multiplication over $\mathbb{G}_T$ and one exponentiation over $\mathbb{G}_T$, respectively. We neglect the computation overhead of the hash operation and multiplication over $\mathbb{Z}_p^*$ as they are not comparable to other operations. Besides, we do not consider the advanced pre-processing or parallelization approaches to expedite the running time, such as the one to accelerate the encryption in Paillier cryptosystem [30], instead we inspect the worst-case computation overhead of our security scheme.

In the authentication phase, it takes $2T_{mul} + 5T_{\exp} + 3T_{par} + 2\widehat{T_{mul}} + 3\widehat{T_{\exp}}$ for each user to generate a signature. On the server side, the HC and LC spend at least $[(2 + |RL|)T_{mul} + 4T_{\exp} + (7 + |RL|)T_{par} + 5\widehat{T_{mul}} + 4\widehat{T_{\exp}}] \times N$ time to validate the authenticity of all participants if no malicious users attempt to double register in the clouds.

In the data collection phase, each user generates two pseudonyms where one is derived from the other one using the Boneh and Franklin's IBE scheme. That means it takes each user $T_{\exp} + T_{par} + \widehat{T_{\exp}}$ time to obtain the pseudonym for HC while the time for generating the pseudonym for LC can be neglected. Later, PHO receives the users' pseudonyms from LC and reverts them using its identity as the decryption key which takes one pairing time per user so it amounts to a total of $T_{par} \times N$ time.

In the data query phase, PHO first obtains the matched pseudonyms from LC. Then, PHO generates a random $N \times N$ invertible matrix $\mathbf{M}$ and two permutation vectors $\boldsymbol{\pi}$ which accounts for the computation complexity of $O\left(N^2\right)$. Afterwards, PHO conducts matrix multiplication (e.g., arithmetic additions/multiplications) and permutation which costs $O\left(KN^2\right)$. The HC, on the other hand, firstly generates secret keys from the Paillier cryptosystem ($O(1)$) and then performs encryption over the health data $\mathbf{H}$ (i.e., $O(N \log n)$). In the stage of **KeyEmd**, PHO performs $O\left(N^2\right)$ multiplications and $O\left(N^2\right)$ additive homomorphic encryptions, thus, its computation complexity is $O(N^2 + N^2 \log n)$. In the end, HC conducts decryption (i.e., $O(N)$) and inner product (i.e., $O(KN)$) to return the vector $\mathbf{CNT}$ to PHO. Furthermore, for notational convenience, we utilize $\overline{T_{PHO}}$ and $\overline{T_{HC}}$ to represent the absolute computation time in aforementioned procedures for

Table I: Communication Overhead

Entity	Process		
	Authentication	Data Collection	Data Query
User	974	320	0
LC	$N \times 974$	$N \times 320$	$N \times 160$
HC	$N \times 974$	$N \times 320$	$N \times 8352 + K \times 64 + N \times K \times 64$
PHO	0	0	$N \times 8512 + K \times 64 + N \times K \times 64$

Table II: Computation Complexity

Entity	Process		
	Authentication	Data Collection	Data Query
User	$2T_{mul} + 5T_{\exp} + 3T_{par} + 2\widehat{T_{mul}} + 3\widehat{T_{\exp}}$	$T_{\exp} + T_{par} + \widehat{T_{\exp}}$	0
LC	$[(2 + \lvert RL\rvert)T_{mul} + 4T_{\exp} + (7 + \lvert RL\rvert)T_{par} + 5\widehat{T_{mul}} + 4\widehat{T_{\exp}}] \times N$	0	0
HC	$[(2 + \lvert RL\rvert)T_{mul} + 4T_{\exp} + (7 + \lvert RL\rvert)T_{par} + 5\widehat{T_{mul}} + 4\widehat{T_{\exp}}] \times N$	0	$\overline{T_{HC}}$
PHO	0	0	$T_{par} \times N + \overline{T_{PHO}}$

PHO and HC, respectively. In so doing, we could summarize the computation complexity of each entity in Table II.

Furthermore, we examine the computational overhead of our design over two real-life datasets on a desktop server. The inhomogeneous Poisson and Bernoulli processes are utilized to test whether an area of disease over-density is indeed statistically significant among others. We generate $R = 999$ replications and set p to be 0.05. The simulation is carried out following two phases, data query and statistic analysis, while the computational overhead for them is measured separately to demonstrate how the incurred overhead due to the security mechanism could impact the overall complexity. The result of running time for data query and statistic analysis is shown in Table III. Under different test distributions (i.e., Poisson and Bernoulli), the detected clusters with the same p-value are the same for two clusters. Specifically, 4 and 20 clusters are found to have outlier disease rates whose p value is less than 0.05 for birth defect and lung cancer datasets, respectively. However, the analysis time using Poisson process is much shorter than that using Bernoulli process. Another insight is that the analysis time using Kulldorf scan statistic is only dependent on the granularity of spatial grids but irrelevant to population. However, the running time of our proposed scheme is actually determined by the number of participants. As we can see, it consumes a significant amount of running time for PHO in the data query phase, which is due to the bilinear pairing operation to decrypt a large number of users' pseudonyms. The HC, on the other hand, takes moderate amount of time using the Paillier cryptosystem during the collaborative computation phase with the PHO. However, the actual PHO/HC server is tremendously powerful (e.g., computer clusters) than our simulation environment; and by further integrating preprocessing or parallelization and considering the participating users are much fewer, the computation overhead is expected to be reduced significantly.

Entity	Dataset			
	Birth Defect in NYS'05		Lung Cancer in NYS'09	
	Poisson	Bernoulli	Poisson	Bernoulli
PHO	2s	7s	140s	188s
	1.72hrs		118.21hrs	
HC	2.82s		216.56s	

Table III: Running Time for Data Query & Analysis

VII. Conclusion

In this paper, we proposed a new paradigm that exploits the multi-cloud platforms to extract users' multi-dimensional data to enable the statistic analysis of infectious diseases. Specifically, we developed a secure protocol to ensure the collected data are not mislead by malicious participants while legitimate users' health and location privacy could be well preserved. Moreover, a novel secure multi-party computation scheme was designed so that the untrusted entities are only allowed to obtain the desired aggregation data for the proposed Kulldorff scan statistic analysis while oblivious of each individual's data records. A rigorous security proof was presented to show our design is both secure and privacy-preserving. Numerical simulations based on real-life datasets also demonstrated the communication and computation overhead of our scheme.

References

[1] A. B. Lawson and K. Kleinman, *Spatial and syndromic surveillance for public health*. John Wiley & Sons, 2005.

[2] "The chief public health officer's report on the state of public health in canada," Online, 2013, http://www.phac-aspc.gc.ca/cphorsphc-respcacsp/2013/infections-eng.php.

[3] L. A. Cole, *The Anthrax Letters: A Bioterrorism Expert Investigates the Attacks that Shocked America*. Skyhorse Publishing Inc., 2009.

[4] G. D. Johnson, "Prospective spatial prediction of infectious disease: experience of new york state (usa) with west nile virus and proposed directions for improved surveillance," *Environmental and ecological statistics*, vol. 15, no. 3, pp. 293–311, 2008.

[5] P. AbdelMalik, M. N. K. Boulos, and R. Jones, "The perceived impact of location privacy: A web-based survey of public health perspectives and requirements in the uk and canada," *BMC Public Health*, vol. 8, no. 1, p. 156, 2008.

[6] "mhealth for development: The opportunity of mobile technology for healthcare in the developing world," Online, 2009, http://www.unfoundation.org/global-issues/technology/mhealth-report.html.

[7] A. Szpiro, B. Johnson, and D. Buckeridge, "Health surveillance and diagnosis for mitigating a bioterror attack," *Lincoln Laboratory J*, vol. 17, no. 1, pp. 101–113, 2007.

[8] S. Van Noort, M. Muehlen, d. A. H. Rebelo, C. Koppeschaar, L. J. Lima, and M. Gomes, "Gripenet: an internet-based system to monitor influenza-like illness uniformly across europe." *Euro surveillance: bulletin européen sur les maladies transmissibles= European communicable disease bulletin*, vol. 12, no. 7, pp. E5–6, 2007.

[9] C. Dalton, D. Durrheim, J. Fejsa, L. Francis, S. Carlson, E. T. d'Espaignet, F. Tuyl *et al.*, "Flutracking: a weekly australian community online survey of influenza-like illness in 2006, 2007 and 2008," *Communicable diseases intelligence quarterly report*, vol. 33, no. 3, p. 316, 2009.

[10] D. Paolotti, C. Gioannini, V. Colizza, and A. Vespignani, "Internet-based monitoring system for influenza-like illness: H1n1 surveillance in italy," in *Proceedings of the 3rd International ICST Conference on Electronic Healthcare for the 21st century. Casablanca*, 2010, pp. 13–15.

[11] O. P. Wójcik, J. S. Brownstein, R. Chunara, and M. A. Johansson, "Public health for the people: participatory infectious disease surveillance in the digital age," *Emerging themes in epidemiology*, vol. 11, no. 1, p. 7, 2014.

[12] N. L. Tilston, K. T. Eames, D. Paolotti, T. Ealden, and W. J. Edmunds, "Internet-based surveillance of influenza-like-illness in the uk during the 2009 h1n1 influenza pandemic," *BMC public health*, vol. 10, no. 1, p. 650, 2010.

[13] M. Debin, C. Turbelin, T. Blanchon, I. Bonmarin, A. Falchi, T. Hanslik, D. Levy-Bruhl, C. Poletto, and V. Colizza, "Evaluating the feasibility and participants' representativeness of an online nationwide surveillance system for influenza in france," *PLoS One*, vol. 8, no. 9, p. e73675, 2013.

[14] E. Uiters, W. Devillé, M. Foets, P. Spreeuwenberg, and P. P. Groenewegen, "Differences between immigrant and non-immigrant groups in the use of primary medical care; a systematic review," *BMC Health Services Research*, vol. 9, no. 1, p. 76, 2009.

[15] 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.

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

[17] M. Kulldorff, F. Mostashari, L. Duczmal, W. Katherine Yih, K. Kleinman, and R. Platt, "Multivariate scan statistics for disease surveillance," *Statistics in medicine*, vol. 26, no. 8, pp. 1824–1833, 2007.

[18] D. Boneh and M. Franklin, "Identity-based encryption from the weil pairing," in *Advances in Cryptology—CRYPTO 2001*. Springer, 2001, pp. 213–229.

[19] D. Boneh and H. Shacham, "Group signatures with verifier-local revocation," in *Proceedings of the 11th ACM conference on Computer and communications security*. ACM, 2004, pp. 168–177.

[20] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, "Private queries in location based services: anonymizers are not necessary," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008, pp. 121–132.

[21] 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.

[22] O. Goldreich, "Secure multi-party computation," *Manuscript. Preliminary version*, pp. 86–97, 1998.

[23] B. Goethals, S. Laur, H. Lipmaa, and T. Mielikäinen, "On private scalar product computation for privacy-preserving data mining," in *International Conference on Information Security and Cryptology*. Springer, 2004, pp. 104–120.

[24] 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.

[25] Y. Zhu, Z. Huang, and T. Takagi, "Secure and controllable k-nn query over encrypted cloud data with key confidentiality," *Journal of Parallel and Distributed Computing*, vol. 89, pp. 1–12, 2016.

[26] 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.

[27] "The java pairing based cryptography library (jpbc)," Online, http://gas.dia.unisa.it/projects/jpbc/.

[28] 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.

[29] "New york state birth defect data," Online, https://www.satscan.org/datasets/nysbirthdefect/index.html.

[30] C. Jost, H. Lam, A. Maximov, and B. J. Smeets, "Encryption performance improvements of the paillier cryptosystem." *IACR Cryptology ePrint Archive*, vol. 2015, p. 864, 2015.