On the Asymptotic Capacity of Information Theoretical Privacy-preserving Epidemiological Data Collection

Jiale Cheng, Nan Liu, and Wei Kang

Abstract

We formulate a new secure distributed computation problem, where a simulation center can require any linear combination of K workers' data through a caching layer consisted of N servers. The workers, servers and the simulation center do not trust each other. For workers, any worker's data is required to be protected from up to E servers; for the server, any more information than the desired linear combination cannot be leaked; and for the simulation center, any single server know nothing about the coefficients of the linear combination. Our goal is to find the optimal download cost which is defined as the size of message uploaded to the simulation center by the servers, to the size of desired linear combination. We proposed a scheme with the optimal download cost when $E < N - 1$. We also prove that when $E \geq N - 1$, the scheme is not feasible.

Index Terms

secure computation, collusion patterns, caching layer

I. Introduction

During the epidemic prevention and control period, strengthening the protection of personal information is not only conducive to safeguarding personal interests, but also better controlling the development of the epidemic. In epidemiological modeling, many recent studies have shown that various models have a good fitting effect on the nature of the epidemic, such as the Bayesian model [1], and the deep learning model including multi-head attention, long short-term memory (LSTM), and convolutional neural network (CNN) [2]. However, the simulation process still can not get rid of the strong dependence on personal data. At the same time, the model adjustment required for a large number of personal data is also a technical problem that needs to be solved urgently. Epidemiological modeling generally requires a collection of different types of information uploaded by users in real-time, but in fact, data collection does not require all the details of users to quantitatively analyze the epidemiological nature. The solution of the contradiction between data analysis and data protection leads us to the theoretical analysis of privacy-preserving epidemiological data collection problem.

In the modeling of epidemiological data collection, a large and changing number of users submit their physical data to an untrusted server at a specified time. Additionally, the data collector who conducts epidemiological modeling retrieves the corresponding data by accessing the server. In real contact graph where physical data is collected by mobile devices, the analytic data of users in some certain regions are essential for epidemiological study to get a proper estimation to the potential public health hazards. Unlike user uploads, these data collectors are only interested in some statistical features contained in the data stored by the server. It is worth noting that users, servers and data collectors do not trust each other, that is, users need to ensure that their data is confidential to the server, and data collectors cannot

J. Cheng and N. Liu are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China (email: {jlcheng, nanliu}@seu.edu.cn). W. Kang is with the School of Information Science and Engineering, Southeast University, Nanjing, China (email: wkang@seu.edu.cn).

This work is partially supported by the National Natural Science Foundation of China under Grants 62071115 and 61971135, and the Research Fund of National Mobile Communications Research Laboratory, Southeast University (No. 2022A03).

know the details of any single user. At the same time, data collectors do not want the server to know the characteristics of user data pairs they are interested in.

In the analysis of communicable diseases including COVID-19, a detailed model with sufficient interaction data is required [3]. Concerning on the data or privacy leakage, Several studies in information theory have focused on the issue when sharing message to untrusted institutions [4]–[12]. In this study, we present a practical framework for privacy-preserving epidemiological data collection problem and analyze the capability that a data collectors can receive the shared data securely and privately, with respect to the number of symbols they need to download.

II. System Model

We formulate the secure privacy-preserving epidemiological data collection problem over a typical distributed secure computation system, in which there are K users, N servers and a data collector, with their respective concerns on data security and privacy. For ease of representation, the entropy and mutual information in the context are all on a basis of a large enough integer $q \in \mathbb{N}_+$. The model is depicted in Fig. 1.

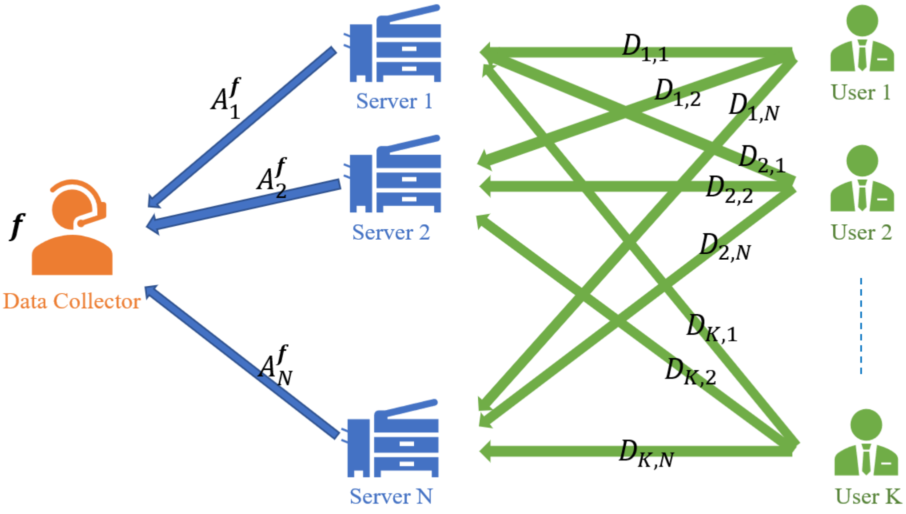

Fig. 1. The Secure Privacy-preserving Epidemiological Data Collection Problem

Assume that user k, $k \in [1:K]$ contains his/her personal message W_k, and the messages of all users are independent and have an equal length of L symbols over a finite field GF_q, i.e.,

$$H(W_{[1:K]}) = \sum_{k=1}^{K} H(W_k), \tag{1}$$

$$H(W_k) = L, \quad \forall k \in [1:K]. \tag{2}$$

The data collection problem contains two phases: in the first phase, which is called *the upload phase*, all K users are required to upload a coded information to each of the N servers, where the uploaded content to the n-th server by the k-th user is denoted as $D_{k,n} \in \mathfrak{D}$. The users would like to keep his/her message secure, more specifically, any up to E servers will learn nothing about the messages uploaded by the K users, i.e.,

$$I(D_{[K],\mathcal{E}}; W_{[K]}) = 0, \quad \forall k \in [1:K], \forall \mathcal{E} \subseteq [1:N], |\mathcal{E}| \leq E, \tag{3}$$

This is called *the privacy constraint of the users against E servers*. In order to achieve this, user k utilizes a privately generated random noise $Z_k \in \mathfrak{Z}_k$, i.e., Z_k is known to only User k, $k \in [1:K]$. Thus, there exists K functions $d_k : GF_q^L \times \mathfrak{Z}_k \to \mathfrak{D}^N, k \in [1:K]$ that $d_k(W_k, Z_k) = \begin{bmatrix} D_{k,1} \ D_{k,2}, \cdots, D_{k,N} \end{bmatrix}^T$, i.e.,

$$H(D_{k,[N]}|W_k, Z_k) = 0, \quad \forall k \in [1:K]. \tag{4}$$

At the beginning of the second phase, which is called *the computation phase*, a data collector would like to compute a statistic of the K messages of the users. In our setting, the statistics, denoted by $W^{\mathbf{f}}$, is taken as a linear combination of all messages $W_{[K]}$ with the coefficient $\mathbf{f} \in GF_q^N$, i.e.,

$$W^{\mathbf{f}} = f(W_{[K]}) = \mathbf{f}^T \begin{bmatrix} W_1 \\ \cdots \\ W_K \end{bmatrix} = \sum_{k=1}^{K} f_k W_k, \tag{5}$$

where the elements of $\mathbf{f}$ are i.i.d. and uniformly distributed on the field GF_q. It is worth noticing that $\mathbf{f}$ is privately generated by the data collector, and it is not known to the users and the servers during the upload phase. Hence, $D_{k,n}$ is not a function of $\mathbf{f}$, $k \in [1:K], n \in [1:N]$.

In order to get the statistics $W^{\mathbf{f}}$, the data collector generates designed queries to each server based on the coefficient $\mathbf{f}$ and a randomnesses $Z' \in \mathcal{Z}$, which is used to protect the security of the coefficient $\mathbf{f}$ against any server. Then we have

$$H(Q_{[1:N]}^{\mathbf{f}}|Z', \mathbf{f}) = 0, \tag{6}$$

where $Q_n^{\mathbf{f}} \in \mathfrak{Q}_n$ denotes the query sent to server n, for all $n \in [N]$, when the coefficient is $\mathbf{f}$. So the queries can be derived from the function $q : GF_q^K \times \mathcal{Z} \to \prod_{n=1}^{N} \mathfrak{Q}_n$ that

$$q(\mathbf{f}, \mathcal{Z}) = \left\{ Q_1^{\mathbf{f}} \ Q_2^{\mathbf{f}} \ \cdots Q_N^{\mathbf{f}} \right\}^T \tag{7}$$

Note that the data collector has no knowledge of the messages of the users, we have

$$I(W_{[1:K]}; Q_{[1:N]}^{\mathbf{f}}, Z', \mathbf{f}) = 0. \tag{8}$$

Upon receiving the query, All N servers are required to calculate the corresponding answers, denoted as $A_{[1:N]}^{\mathbf{f}}$, to the data collector. More specifically, the answer generated by Server n, i.e., $A_n^{\mathbf{f}} \in \mathfrak{A}_n$, is a deterministic function of the stored content of Server n, i.e., $D_{[1:K],n}$, and the query it received, i.e., $Q_n^{\mathbf{f}}$. In other words, we have the function

$$a_n^{\mathbf{f}} : \mathfrak{Q}_n \times \mathfrak{D}^K \to \mathfrak{A}_n, \ A_n^{\mathbf{f}} = a_n^{\mathbf{f}}(Q_n^{\mathbf{f}}, \begin{bmatrix} D_{1,n} \ D_{2,n}, \cdots, D_{K,n} \end{bmatrix}^T), \quad n \in [1:N], \tag{9}$$

it could be written as:

$$H(A_n^{\mathbf{f}}|Q_n^{\mathbf{f}}, D_{[1:K],n}) = 0, \quad \forall n \in [1:N]. \tag{10}$$

We would like to design the queries that meet the following 2 constraints. The first requires the data collector is able to reconstruct the desired statistics from all the answers from the N servers, which we call *the decodability constraint*. Let ϕ be the reconstruction function of data collector, we have

$$\begin{aligned} &\phi : \prod_{n=1}^{N} \mathfrak{A}_n \times \prod_{n=1}^{N} \mathfrak{Q}_n \times GF_q^K \times \mathcal{Z} \to GF_q^L, \\ &\hat{W}^{\mathbf{f}} = \phi(A_{[1:N]}^{\mathbf{f}}, Q_{[1:N]}^{\mathbf{f}}, \mathbf{f}, Z'). \end{aligned} \tag{11}$$

The probability of error (PoE) in (11) is defined by

$$P_e = Pr\{\hat{W}^{\mathbf{f}} \neq W^{\mathbf{f}}\}. \tag{12}$$

According to the Fano's inequality, the reconstruction constraint is equal to

$$\frac{1}{L}H(W^{\mathbf{f}}|A^{\mathbf{f}}_{[1:N]},Q^{\mathbf{f}}_{[1:N]},\mathbf{f},Z')=o(L). \tag{13}$$

For the second restriction, any single server must learns nothing about the messages of the users, which we call *the privacy constraint of the users against the data collector*, i.e.,

$$I(W_{[1:K]};A^{\mathbf{f}}_{[1:N]},Q^{\mathbf{f}}_{[1:N]},\mathbf{f},Z'|W^{\mathbf{f}})=0. \tag{14}$$

We further assume that the servers are curious about the coefficients of the statistics, i.e., $\mathbf{f}$. To protect the privacy of the data collector, we require that the coefficient vector $\mathbf{f}$ is not leaked to any single server, i.e.,

$$I(\mathbf{f};A^{\mathbf{f}}_{n},Q^{\mathbf{f}}_{n},W_{[1:K]})=0, \quad \forall n\in[1:N]. \tag{15}$$

This is called *the privacy constraint of the data collector against non-colluding servers*.

For any scheme that satisfies the above reconstruction constraint, i.e., (13), and the privacy constraints, i.e., the privacy constraint of the users against E servers (3), the privacy constraint of the users against the data collector (14), and the privacy constraint of the data collector against the non-colluding servers (15), its communication rate is characterized by the number of symbols the data collector downloads per decoded symbol, i.e.,

$$R:=\frac{H(W^{\mathbf{f}}|\mathbf{f})}{\sum_{n=1}^{N}H(A^{\mathbf{f}}_{n}|\mathbf{f})}=\frac{L}{\sum_{n=1}^{N}H(A^{\mathbf{f}}_{n})}, \tag{16}$$

note that R is not a function of $\mathbf{f}$ due to (15).

A rate R is said to be (ϵ-error) achievable if there exists a sequence of schemes with their communication rate less than or equal to R where the probability of error P_e goes to zero when $L\rightarrow\infty$. The ϵ-error capacity of this random secure aggregation problem is defined as the supremum of all ϵ-error achievable rates, i.e., $C:=\inf R$, where the infimum is over all possible ϵ-error achievable schemes.

III. Main Result

Theorem 1 *When $K\rightarrow\infty$, the asymptotic capacity of secure private epidemiological statistics under E-security and T-colluding servers is formulated by*

$$\lim_{K\rightarrow\infty,L\rightarrow\infty}C=\begin{cases}\frac{N-E-1}{N}, & \text{if } E<N-1\\ 0, & \text{otherwise}\end{cases}, \tag{17}$$

The converse proof of Theorem 1 will be given in Section IV, and the achievability for a certain cases of finite $K\in\mathbb{N}_+$ will be given in Section V. Noticing that when $K\rightarrow\infty$, the schemes in Section V can be achievable by sending multiple rounds of queries in the same strategy, the rate of achievability and converse will meet when K goes infinite.

We make some remarks here regarding Theorem 1.

Remark 1 *We notice that in practical settings, the upload phase and the computation phase do not always occur at the same time. For example, the users are required to upload their epidemiological data on a regular basis, while the data collector may start their queries to a certain statistics at a relatively random time. Due to the dynamic topology of the servers, the numbers of connected(colluding) servers may be different during the uploading phase and the computation phase. Our work only concentrates on the case where the servers are non-colluding in the computation phase, considering that the servers may be more interested in the epidemiological data. Specifically, when we set $E=1$, i.e., the servers will always be non-colluding, the results has a similar form as the T-SPIR problem [13] when $T=2$. The problems of collusion of servers in the computation phase is still an open problem.*

Remark 2 *When the number of message K is a finite integer, the achievability and converse in our work will not meet. From our work, the converse of finite K seems to be depend on K, while the scheme we construct is irrelevant to K in order to protect the privacy of the users against the data collector. How to diminish or eliminate the gap for some cases when K is finite is still an open problem.*

IV. Proof of Theorem 1: Converse when $E < N - 1$

In this section, we prove the converse part of Theorem 1. Firstly, we propose two lemmas related to the symmetric constraint:

Lemma 1 *Assume that $\mathbf{f}$ and $\mathbf{f}'$ are linear-independent vectors of GF_q^N,for any $\mathcal{E} \subseteq [1:N], |\mathcal{E}| = E$, we have the following equalities:*

$$H(A_{\mathcal{E}}^{\mathbf{f}}|Q_{\mathcal{E}}^{\mathbf{f}}) = H(A_{\mathcal{E}}^{\mathbf{f}'}|Q_{\mathcal{E}}^{\mathbf{f}'}) \tag{18}$$

$$H(A_{\mathcal{E}}^{\mathbf{f}}|Q_{\mathcal{E}}^{\mathbf{f}}, W^{\mathbf{f}}) = H(A_{\mathcal{E}}^{\mathbf{f}'}|Q_{\mathcal{E}}^{\mathbf{f}'}, W^{\mathbf{f}}) \tag{19}$$

Lemma 2 *For any $\mathcal{E} \subseteq [1:N], |\mathcal{E}| = E$,*

$$H(A_{\mathcal{E}}^{\mathbf{f}}|Q_{\mathcal{E}}^{\mathbf{f}}, W^{f}) = H(A_{\mathcal{E}}^{\mathbf{f}}|Q_{\mathcal{E}}^{\mathbf{f}}) \tag{20}$$

Lemma 3

$$H(W^{\mathbf{f}}) \leq H(A_{[1:N]}^{\mathbf{f}}|Q_{[1:N]}^{\mathbf{f}}) - H(A_{\mathcal{E}}^{\mathbf{f}}|Q_{\mathcal{E}}^{\mathbf{f}}) \tag{21}$$

The proofs use the same technique in [14], and we omit the details here.

The following lemma shows that we can split the answers into two parts, one from E servers that cannot decode the database and the other from $N - E$ servers:

Lemma 4 *For any $\mathbf{f} \in GF_q^K$ and $\mathcal{E} \in [1:N]$, $|\mathcal{E}| = E$, we have*

$$\begin{aligned}&\left(1 - \frac{E}{N}\right) H(A_{[1:N]}^{\mathbf{f}}|Q_{[1:N]}^{\mathbf{f}})\\ \geq& L + H(A_{[1:N]/\mathcal{E}}^{\mathbf{f}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, Q_{[1:N]}^{\mathbf{f}}) - o(L)\end{aligned} \tag{22}$$

Proof:

$$\begin{aligned}
&H(A_{[1:N]}^{\mathbf{f}}|Q_{[1:N]}^{\mathbf{f}}) && (23)\\
=&H(W^{\mathbf{f}}|Q_{[1:N]}^{\mathbf{f}}) + H(A_{[1:N]}^{\mathbf{f}}|W^{\mathbf{f}}, Q_{[1:N]}^{\mathbf{f}})\\
&- H(W^{\mathbf{f}}|A_{[1:N]}^{\mathbf{f}}, Q_{[1:N]}^{\mathbf{f}}) && (24)\\
=&L + H(A_{[1:N]}^{\mathbf{f}}|W^{\mathbf{f}}, Q_{[1:N]}^{\mathbf{f}}) - o(L) && (25)\\
=&L + H(A_{\mathcal{E}}^{\mathbf{f}}|W^{\mathbf{f}}, Q_{[1:N]}^{\mathbf{f}})\\
&+ H(A_{[1:N]/\mathcal{E}}^{\mathbf{f}}|W^{\mathbf{f}}, A_{\mathcal{E}}^{\mathbf{f}}, Q_{[1:N]}^{\mathbf{f}}) - o(L) && (26)\\
=&L + H(A_{\mathcal{E}}^{\mathbf{f}}|Q_{[1:N]}^{\mathbf{f}})\\
&+ H(A_{[1:N]/\mathcal{E}}^{\mathbf{f}}|W^{\mathbf{f}}, A_{\mathcal{E}}^{\mathbf{f}}, Q_{[1:N]}^{\mathbf{f}}) - o(L) && (27)\\
\geq&L + H(A_{\mathcal{E}}^{\mathbf{f}}|Q_{[1:N]}^{\mathbf{f}})\\
&+ H(A_{[1:N]/\mathcal{E}}^{\mathbf{f}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, Q_{[1:N]}^{\mathbf{f}}) - o(L) && (28)\\
\geq&L + \frac{E}{N} H(A_{[1:N]}^{\mathbf{f}}|Q_{[1:N]}^{\mathbf{f}})\\
&+ H(A_{[1:N]/\mathcal{E}}^{\mathbf{f}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, Q_{[1:N]}^{\mathbf{f}}) - o(L) && (29)
\end{aligned}$$

■

where the last inequality follows from the Han's inequality,

$$\sum_{\mathcal{E}\subseteq[1:N],|\mathcal{E}|=E} H(A^{\mathbf{f}}_{\mathcal{E}}|Q^{\mathbf{f}}_{\mathcal{E}}) \geq \frac{E}{N}\binom{N}{E} H(A^{\mathbf{f}}_{[1:N]}|Q^{\mathbf{f}}_{[1:N]}) \tag{30}$$

To give an upper bound of the second term of (22), we have the following iteration relation.

Lemma 5 *Let* $\mathbf{f}, \mathbf{f}' \in GF_q^K$ *be linear independent vectors, and* $\mathcal{E} \in [1:N]$, $|\mathcal{E}| = E$, *we have*

$$\begin{aligned}&H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}}_{[1:N]})\\ \geq&\frac{1}{N-E}\left(L + H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, W^{\mathbf{f}'}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}'}_{[1:N]})\right)\end{aligned} \tag{31}$$

Proof:

$$\begin{aligned}&(N-E)H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}}_{[1:N]})\\ \geq&\sum_{n\in[1:N]/\mathcal{E}} H(A^{\mathbf{f}}_{n}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}}_{[1:N]}) &(32)\\ =&\sum_{n\in[1:N]/\mathcal{E}} H(A^{\mathbf{f}'}_{n}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}'}_{[1:N]}) &(33)\\ \geq&H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}'}_{[1:N]}) &(34)\\ =&H(W^{\mathbf{f}'}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}'}_{[1:N]})\\ &+H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}'}, W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}'}_{[1:N]}) &(35)\\ =&L + H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}'}, W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}'}_{[1:N]}) &(36)\end{aligned}$$

■

the last equation holds because $W^{\mathbf{f}'}$ is independent from the queries and randomness, the security constraint and that $\mathbf{f}, \mathbf{f}' \in GF_q^K$ are linear independent vectors.

Then we can get the upper bound of the asymptotic download when L and K goes infinity:

$$\begin{aligned}&\lim_{K\to\infty,L\to\infty}\left(1-\frac{E}{N}\right)H(A^{\mathbf{f}}_{[1:N]}|Q^{\mathbf{f}}_{[1:N]})\\ \geq&\lim_{K\to\infty,L\to\infty} H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}}_{[1:N]}) - o(L) + L &(37)\\ \geq&\lim_{K\to\infty} H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}}_{[1:N]}) + L &(38)\\ \geq&\lim_{K\to\infty}\frac{1}{N-E}\left(L + H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}'}_{[1:N]})\right) + L &(39)\\ =&\left(\sum_{k=0}^{\infty}\frac{1}{(N-E)^k}\right)L &(40)\end{aligned}$$

and thus we get the upper bound when $E < N-1$:

$$\begin{aligned}\lim_{K\to\infty,L\to\infty} C_\rho =&\frac{L}{H(A^{\mathbf{f}}_{[1:N]})} &(41)\\ \leq&\frac{L}{H(A^{\mathbf{f}}_{[1:N]}|H(Q^{\mathbf{f}}_{[1:N]})} &(42)\\ \leq&\frac{1-\frac{E}{N}}{\sum_{k=0}^{\infty}\frac{1}{(N-E)^k}} &(43)\end{aligned}$$

$$=\frac{N-E-1}{N}. \tag{44}$$

V. Proof of Theorem 1: Achievability when $E < N-1$

We give a cross subspace alignment (CSA) scheme based on the coding on interference in the computation phase to reach the asymptotic capacity [15] for any N, K. Throughout the scheme, we choose $L = N-E-1$, and we use the notation $\Delta_n = \prod_{i=1}^{L}(i+\alpha_n)$ for $n \in [1:N]$.

First, we specify the encoding functions $\{d_k\}_{k\in[1:K]}$ in the upload phase. Let $W_k^i \in GF_q$ be the i-th symbol of each W_k, $k \in [1:K]$ and $W^i = (W_1^i, \cdots, W_K^i)$. Assume that $\alpha_n \in \{\alpha \in GF_q : \alpha + i \neq 0, i \in [1:L]\}$ for all $n \in [1:N]$ are globally shared variables, known with the users, servers and the data collector. In order to protect the privacy of the users against the servers, we introduce the noises $Z_{le} \in GF_q^{1\times K}, l \in [1:L], e \in [1:E]$ shared by all of the users. We have $D_{k,n} \in \mathfrak{D} = GF_q^{1\times L}$ and $Z = \{Z_{le}\}_{l\in[1:L],e\in[1:E]} \in \mathfrak{Z} = GF_q^{LE\times K}$ for all $k \in [1:K], n \in [1:N]$ in the proposed shceme, and the storage of each server n is designed as the additional noisy data, i.e.,

$$D_{k,n} = \mathbf{e}_n^T d_k(W_k, Z) = \begin{bmatrix} W_k^1 + \sum_{e=1}^{E}(1+\alpha_n)^e Z_{1e}(k) \\ \cdots \\ W_k^L + \sum_{e=1}^{E}(L+\alpha_n)^e Z_{Le}(k) \end{bmatrix}^T \forall k \in [1:K], n \in [1:N], \tag{45}$$

where $\mathbf{e}_n \in \{0,1\}^N$ is the unitary vector with only the n-th entry being 1. To simplify the notation, we write them in a vector. Let $D_{n,k}^i$ denote the i-th symbol of $D_{n,k}$, and $D_n \in GF_q^{1\times KL}$ denotes the storage of server n, we have

$$D_n = (D_{n,1}^1, \cdots, D_{n,K}^1, D_{n,1}^2, \cdots, D_{n,K}^2, \cdots, D_{n,1}^L, \cdots, D_{n,K}^L) \tag{46}$$

$$= \begin{bmatrix} W^1 + \sum_{e=1}^{E}(1+\alpha_n)^e Z_{1e} \\ \cdots \\ W^L + \sum_{e=1}^{E}(L+\alpha_n)^e Z_{Le}. \end{bmatrix}^T \tag{47}$$

In the computation phase, the query to database n is determined by the coefficient $\mathbf{f}$ and the randomness from data collector Z'. In our scheme, we have $\mathcal{Z} = GF_q^{1\times K}$ and $\mathfrak{Q}_n = GF_q^{KL}$ for all $n \in [1:N]$, i.e., the query is formulated as an $LK \times 1$ matrix. The function q is characterized by

$$Q_n^{\mathbf{f}} = \mathbf{e}_n^T q(\mathbf{f}, \mathcal{Z}) = \begin{bmatrix} \frac{\Delta_n}{1+\alpha_n}(\mathbf{f} + (1+\alpha_n)(Z_1')^T); \\ \cdots; \\ \frac{\Delta_n}{L+\alpha_n}(\mathbf{f} + (L+\alpha_n)(Z_L')^T). \end{bmatrix} \tag{48}$$

For any server $n \in [1:N]$, the answer to data collector $A_n^{\mathbf{f}} \in \mathfrak{A}_n = GF_q$ is calculated by multiplying the storage and the query, i.e.,

$$A_n^{\mathbf{f}} = D_n \cdot Q_n^{\mathbf{f}} \tag{49}$$

$$\begin{aligned} =& \left(W^1 + \sum_{e=1}^{E}(1+\alpha_n)^e Z_{1e}\right) \cdot \left(\frac{\Delta_n}{1+\alpha_n}(\mathbf{f} + (1+\alpha_n)(Z_1')^T)\right) \\ &+ \cdots + \left(W^L + \sum_{e=1}^{E}(L+\alpha_n)^e Z_{Le}\right) \\ &\cdot \left(\frac{\Delta_n}{L+\alpha_n}(\mathbf{f} + (L+\alpha_n)(Z_L')^T)\right), \forall n \in [1:N]. \end{aligned} \tag{50}$$

Now we take a closer look at the expansion of terms in sum (50), and divide them into message terms, interference terms and noisy terms.

Firstly, there are desired signals $\frac{\Delta_n}{l+\alpha_n} W^l \cdot \mathbf{f}$ for all $l \in [1:L]$, The L symbols are desired symbols that needs to be decoded, as

$$W^{\mathbf{f}} = W^T \cdot \mathbf{f} = \begin{bmatrix} W^1 \cdot \mathbf{f} \\ \cdots \\ W^L \cdot \mathbf{f} \end{bmatrix}, \tag{51}$$

and that $\frac{\Delta_n}{1+\alpha_n}$ is globally known.

In contrast, the interference terms in the answer $A_n^{\mathbf{f}}$ with respect to $W^l, l \in [1:L]$ is represented by a non-negative power of $(l+\alpha_n)$, so it can be denoted by a polynomial $\Delta_n \sum_{e=0}^{E} (l+\alpha_n)^e I_{le}$, where I_{le} is some sums of $W^l \cdot (Z_l')^T$, $Z_{le} \cdot \mathbf{f}$ and $Z_{le}\mathbf{f} \cdot (Z_l')^T$. So we can re-write $A_n^{\mathbf{f}}$ according to the ascending power of $(l+\alpha_n)$ (and thus of α_n) for every $l \in [1:L]$, i.e.,

$$\begin{aligned} A_n^{\mathbf{f}} =& \Delta_n \sum_{l=1}^{L} \left(W^l \cdot \mathbf{f} (l+\alpha_n)^{-1} + \sum_{e=0}^{E} I_{le} (l+\alpha_n)^e \right) && (52) \\ =& \Delta_n \left(\sum_{l=1}^{L} W^l \cdot \mathbf{f} (l+\alpha_n)^{-1} + \sum_{e=0}^{E} I_e' \alpha_n^e \right), \forall n \in [1:N]. && (53) \end{aligned}$$

We write the answers the user received in a matrix form:

$$\begin{bmatrix} \frac{A_1^{\mathbf{f}}}{\Delta_1} \\ \frac{A_2^{\mathbf{f}}}{\Delta_2} \\ \cdots \\ \frac{A_N^{\mathbf{f}}}{\Delta_N} \end{bmatrix} = \begin{bmatrix} \frac{1}{1+\alpha_1} & \cdots & \frac{1}{L+\alpha_1} & 1 & \alpha_1 & \cdots & \alpha_1^E \\ \frac{1}{1+\alpha_2} & \cdots & \frac{1}{L+\alpha_2} & 1 & \alpha_2 & \cdots & \alpha_2^E \\ & & & \cdots & & & \\ \frac{1}{1+\alpha_N} & \cdots & \frac{1}{L+\alpha_N} & 1 & \alpha_N & \cdots & \alpha_N^E \end{bmatrix} \cdot \begin{bmatrix} W^1 \cdot \mathbf{f} \\ \cdots \\ W^L \cdot \mathbf{f} \\ I_0' \\ \cdots \\ I_E' \end{bmatrix}, \tag{54}$$

where we denote the medium matrix by M_E.

Now we prove that this scheme satisfy the correctness, privacy and security against servers and data collector.

The correctness to data collector is satisfied because the middle transformation matrix is a $N \times N$ full-rank matrix, so the data collector can decode the sum $W^{\mathbf{f}}$:

$$\begin{aligned} W^{\mathbf{f}} =& \begin{bmatrix} W^1 \cdot \mathbf{f} \\ \cdots \\ W^L \cdot \mathbf{f} \end{bmatrix} && (55) \\ =& \begin{bmatrix} \frac{1}{1+\alpha_1} & \cdots & \frac{1}{L+\alpha_1} & 1 & \alpha_1 & \cdots & \alpha_1^E \\ \frac{1}{1+\alpha_2} & \cdots & \frac{1}{L+\alpha_2} & 1 & \alpha_2 & \cdots & \alpha_2^E \\ & & & \cdots & & & \\ \frac{1}{1+\alpha_N} & \cdots & \frac{1}{L+\alpha_N} & 1 & \alpha_N & \cdots & \alpha_N^E \end{bmatrix}^{-1} \cdot \begin{bmatrix} \frac{A_1^{\mathbf{f}}}{\Delta_1} \\ \frac{A_2^{\mathbf{f}}}{\Delta_2} \\ \cdots \\ \frac{A_N^{\mathbf{f}}}{\Delta_N} \end{bmatrix} && (56) \\ & \{[1 : N-E-1], :\}. && (57) \end{aligned}$$

The security against servers is protected due to the sharing strategy of users. In (45), we know that the k-th user share its l-th symbol to the n-th server in a form

$$D_{n,k}^l = W_k^l + \sum_{e=1}^{E} (1+\alpha_n)^e Z_{le}(k), \tag{58}$$

where $D_{n,k}^l$ denotes the storages in server n that W_k^l shares. The security need to guarantee that any E out of N servers do not know W_k^l for any $k \in [1:K]$ and $l \in [1:L]$. To prove this, we write the storages of E servers with respect to what W_k^l shares in a matrix form. Here we choose the servers to be in $[1:E]$, w.l.o.g., and

$$\begin{aligned}\begin{bmatrix} D^l_{1,k} \\ \cdots \\ D^l_{E,k} \end{bmatrix} = &\begin{bmatrix} W^l_k \\ \cdots \\ W^l_k \end{bmatrix} \\ &+ \begin{bmatrix} l+\alpha_1 \ (l+\alpha_1)^2 \ \cdots \ (l+\alpha_1)^E \\ \cdots \\ l+\alpha_E \ (l+\alpha_1)^E \ \cdots \ (l+\alpha_E)^E \end{bmatrix} \cdot \begin{bmatrix} Z_{l1}(k) \\ \cdots \\ Z_{lE}(k) \end{bmatrix}. \end{aligned} \tag{59}$$

The Vandermonde matrix in (59), denoted by V_E is invertible for distinct $\{1+\alpha_e : \alpha_e \in GF_q, e \in [1:E]\}$, so the second term of (59) contains E symbols that are linearly independent, so we have

$$\begin{aligned} &I(D_{k,\mathcal{E}}; W_k) && (60)\\ =&\sum_{l=1}^{L} I(D^l_{k,\mathcal{E}}; W^l_k) && (61)\\ =&\sum_{l=1}^{L} I\left(W^l_k; W^l_k \mathbf{1}_E + V_E \cdot \begin{bmatrix} Z_{l1}(k) \\ \cdots \\ Z_{lE}(k) \end{bmatrix} \right) && (62)\\ =&\sum_{l=1}^{L} I\left(W^l_k; Z_{l1}(k), \cdots, Z_{lE}(k) \right) && (63)\\ =&0, \quad \forall k \in [1:K], n \in [1:N]. && (64) \end{aligned}$$

To prove the privacy constraint of the data collector against non-colluding servers, we notice that the query to each server is composed of the desired coefficient $\mathbf{f}$ and independent additional noise Z'_l, $l \in [1:L]$. Thus, privacy is protected, i.e.,

$$\begin{aligned} &I(Q^{\mathbf{f}}_n, A^{\mathbf{f}}_n, W_{[1:K]}; \mathbf{f}) && (65)\\ \leq& I(Q^{\mathbf{f}}_n, W_{[1:K]}, Z; \mathbf{f}) && (66)\\ =& I(Q^{\mathbf{f}}_n; \mathbf{f} | W_{[1:K]}, Z) && (67)\\ \leq& H(Q^{\mathbf{f}}_n) - H\left(\begin{bmatrix} \mathbf{f} + (1+\alpha_n)(Z'_1) \\ \cdots \\ \mathbf{f} + (L+\alpha_n)(Z'_L) \end{bmatrix} \middle| \mathbf{f}, W_{[1:K]}, Z \right) && (68)\\ =& H(Q^{\mathbf{f}}_n) - H(Z'_1, \cdots, Z'_L) && (69)\\ =& L - L \\ =& 0, \end{aligned}$$

where (66) is from (50) and (4), and (67) is from (14).

To prove the security against data collector, we construct a base containing the desired $\mathbf{f}$ of GF_q^K and the vectors in the basis is denoted by $\{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_K\}$ where $\mathbf{f}_1 = \mathbf{f}$. We then have

$$\begin{aligned} &I\left(W_{[1:K]}; A^{\mathbf{f}}_{[1:N]} | W^{\mathbf{f}} \right) && (70)\\ =&\sum_{l=1}^{L} I\left(W^l; A^{\mathbf{f}}_{[1:N]} | W^{[1:l]}, W^{\mathbf{f}} \right) && (71)\\ \leq&\sum_{l=1}^{L} I\left(W^l; A^{\mathbf{f}}_{[1:N]} | W^{[1:L]/\{l\}}, Z, W^{\mathbf{f}} \right) && (72) \end{aligned}$$

$$
\begin{aligned}
=&\sum_{l=1}^{L} I\left(\left(W^{l}+\sum_{e=1}^{E}(l+\alpha_n)^{e}Z_{le}\right)\cdot\right.\\
&\qquad\left.\left(\frac{\Delta_n}{1+\alpha_n}(\mathbf{f}+(l+\alpha_n)(Z_l')^{T})\right)_{n\in[1:N]};W^{l}|W^{[1:L]/\{l\}},Z,W^{\mathbf{f}}\right) &\text{(73)}\\
=&\sum_{l=1}^{L} I\left(\left(W^{l}(Z_l')^{T}+\sum_{e=1}^{E}(l+\alpha_n)^{e}Z_{le}(Z_l')^{T}\right)_{n\in[1:N]};W^{l}|W^{[1:L]/\{l\}},W^{\mathbf{f}}\right) &\text{(74)}\\
\leq&\sum_{l=1}^{L} I\left(\left(W^{l}(Z_l')^{T},Z_{le}(Z_l')^{T}\right)_{n\in[1:N],e\in[1:E]};W^{l}|W^{[1:L]/\{l\}},W^{\mathbf{f}}\right) &\text{(75)}\\
=&0, &\text{(76)}
\end{aligned}
$$

where (72) holds because $(W^{[1:L]/\{l\}},Z)$ is independent with W^{l}, (73) holds because except for the term containing W^{l}, all terms in (50) are given, so deducting them will not change the mutual information. (74) is because Δ_n is a constant. Thus, we can prove that the scheme satisfies all the constraints. As any server answer is a symbol from GF_q, the rate in the proposed scheme is

$$
R=\frac{L}{H(A_n^{\mathbf{f}})}=\frac{N-E-1}{N} \tag{77}
$$

We notice that the achievable rate meets the asymptotic upper bound for any $K\in\mathbb{N}_+$, so the scheme is then proved to be asymptotically optimal, by letting $K\to\infty$.

VI. Proof of Theorem 1: Converse when $E\geq N-1$

We know that when $N=E$, the correctness and security constraints contradicts each other and any scheme is not feasible to the problem, so we focus on the case when $N=E+1$.

For converse, the inequality (22) also holds, as the inequality in (32) becomes an equality, and we have

$$
\begin{aligned}
&H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}},D_{[1:K],\mathcal{E}},S,Q^{\mathbf{f}}_{[1:N]})\\
\geq&\left(L+H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}},W^{\mathbf{f}'},D_{[1:K],\mathcal{E}},S,Q^{\mathbf{f}'}_{[1:N]})\right) &\text{(78)}
\end{aligned}
$$

so we have

$$
\begin{aligned}
&\left(1-\frac{E}{N}\right)H(A^{\mathbf{f}}_{[1:N]}|Q^{\mathbf{f}}_{[1:N]})\\
\geq&H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}},D_{[1:K],\mathcal{E}},S,Q^{\mathbf{f}}_{[1:N]})-o(L)+L &\text{(79)}\\
\geq&H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}},D_{[1:K],\mathcal{E}},S,Q^{\mathbf{f}}_{[1:N]})+L &\text{(80)}\\
\geq&\left(L+H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}},D_{[1:K],\mathcal{E}},S,Q^{\mathbf{f}'}_{[1:N]})\right)+L &\text{(81)}\\
=&\sum_{k=0}^{K}L+H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}}|W_{[1:N]},D_{[1:K],\mathcal{E}},S,Q^{\mathbf{f}'\prime}_{[1:N]}) &\text{(82)}\\
=&KL &\text{(83)}
\end{aligned}
$$

and as the download is tend to go infinite when$K\to\infty$, the asymptotic capacity is zero:

$$
\begin{aligned}
\lim_{K\to\infty,L\to\infty}C_\rho=&\lim_{K\to\infty,L\to\infty}\frac{L}{H(A^{\mathbf{f}}_{[1:N]})} &\text{(84)}\\
\leq&\lim_{K\to\infty,L\to\infty}\frac{L}{H(A^{\mathbf{f}}_{[1:N]}|Q^{\mathbf{f}}_{[1:N]})} &\text{(85)}
\end{aligned}
$$

$$\leq \lim_{K\to\infty, L\to\infty} \frac{1-\frac{E}{N}}{KL} \tag{86}$$

$$=0 \tag{87}$$

Therefore, it is unfeasible to construct a scheme that has a positive asymptotic capacity when $K \to \infty$.

VII. Conclusion

We have modeled and found the asymptotical capacity of the privacy-preserving epidemiological data collection problem. We show that when there are more than 1 remaining servers that do not collude with other servers to decode the users' data, the asymptotical capacity exists. The results in this work shows a similar capacity form with symmetric private information retrieval.

References

[1] Sean C. Anderson, Andrew M. Edwards, Madi Yerlanov, Nicola Mulberry, Jessica E. Stockdale, Sarafa A. Iyaniwura, Rebeca C. Falcão, Michael C. Otterstatter, Michael A. Irvine, Naveed Zafar Janjua, Daniel Coombs, and Caroline Colijn. Quantifying the impact of covid-19 control measures using a bayesian model of physical distancing. *PLoS Computational Biology*, 16, 2020.

[2] Hossein Abbasimehr and Reza Paki. Prediction of covid-19 confirmed cases combining deep learning methods and bayesian optimization. *Chaos, Solitons & Fractals*, 142:110511, 2021.

[3] Daniel Gnther, Marco Holz, Benjamin Judkewitz, Helen Mllering, Benny Pinkas, and Thomas Schneider. Pem: Privacy-preserving epidemiological modeling. Cryptology ePrint Archive, Report 2020/1546, 2020.

[4] Kai Wan, Hua Sun, Mingyue Ji, and Giuseppe Caire. Distributed linearly separable computation. *IEEE Transactions on Information Theory*, pages 1–1, 2021.

[5] Kai Wan, Hua Sun, Mingyue Ji, and Giuseppe Caire. On the tradeoff between computation and communication costs for distributed linearly separable computation. *IEEE Transactions on Communications*, 69(11):7390–7405, 2021.

[6] Kai Wan, Hua Sun, Mingyue Ji, and Giuseppe Caire. On secure distributed linearly separable computation. *IEEE Journal on Selected Areas in Communications*, 40(3):912–926, 2022.

[7] Zhen Chen, Zhuqing Jia, Zhiying Wang, and Syed A Jafar. Gcsa codes with noise alignment for secure coded multi-party batch matrix multiplication. *arXiv: Information Theory*, 2020.

[8] Yizhou Zhao and Hua Sun. Information theoretic secure aggregation with user dropouts. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1124–1129, 2021.

[9] Wei-Ting Chang and Ravi Tandon. On the capacity of secure distributed matrix multiplication. In *2018 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2018.

[10] Burak Hasrcolu, Jess Gmez-Vilardeb, and Deniz Gndz. Bivariate polynomial codes for secure distributed matrix multiplication. *IEEE Journal on Selected Areas in Communications*, 40(3):955–967, 2022.

[11] Netanel Raviv, Itzhak Tamo, Rashish Tandon, and Alexandros G. Dimakis. Gradient coding from cyclic mds codes and expander graphs. *IEEE Transactions on Information Theory*, 66(12):7475–7489, 2020.

[12] Jiale Cheng, Nan Liu, Wei Kang, and Yang Li. The capacity of symmetric private information retrieval under arbitrary collusion and eavesdropping patterns. *IEEE Transactions on Information Forensics and Security*, 17:3037–3050, 2022.

[13] Qiwen Wang and Mikael Skoglund. On PIR and symmetric PIR from colluding databases with adversaries and eavesdroppers. *IEEE Transactions on Information Theory*, 65(5):3183–3197, May 2019.

[14] Qiwen Wang, Hua Sun, and Mikael Skoglund. The capacity of private information retrieval with eavesdroppers. *IEEE Transactions on Information Theory*, 65(5):3198–3214, May 2018.

[15] Zhuqing Jia, Hua Sun, and Syed Ali Jafar. Cross subspace alignment and the asymptotic capacity of x -secure t -private information retrieval. *IEEE Transactions on Information Theory*, 65(9):5783–5798, 2019.