entropy

MDPI

Article

On the Asymptotic Capacity of Information-Theoretic Privacy-Preserving Epidemiological Data Collection

Jiale Cheng [1], Nan Liu [1,*] and Wei Kang [2]

[1] National Mobile Communications Research Laboratory, Southeast University, Nanjing 211189, China
[2] School of Information Science and Engineering, Southeast University, Nanjing 211189, China
* Correspondence: nanliu@seu.edu.cn

Abstract: The paradigm-shifting developments of cryptography and information theory have focused on the privacy of data-sharing systems, such as epidemiological studies, where agencies are collecting far more personal data than they need, causing intrusions on patients' privacy. To study the capability of the data collection while protecting privacy from an information theory perspective, we formulate a new distributed multiparty computation problem called privacy-preserving epidemiological data collection. In our setting, a data collector requires a linear combination of K users' data through a storage system consisting of N servers. Privacy needs to be protected when the users, servers, and data collector do not trust each other. For the users, any data are required to be protected from up to E colluding servers; for the servers, any more information than the desired linear combination cannot be leaked to the data collector; and for the data collector, any single server can not know anything about the coefficients of the linear combination. Our goal is to find the optimal collection rate, which is defined as the ratio of the size of the user's message to the total size of downloads from N servers to the data collector. For achievability, we propose an asymptotic capacity-achieving scheme when $E < N - 1$, by applying the cross-subspace alignment method to our construction; for the converse, we proved an upper bound of the asymptotic rate for all achievable schemes when $E < N - 1$. Additionally, we show that a positive asymptotic capacity is not possible when $E \geq N - 1$. The results of the achievability and converse meet when the number of users goes to infinity, yielding the asymptotic capacity. Our work broadens current researches on data privacy in information theory and gives the best achievable asymptotic performance that any epidemiological data collector can obtain.

Keywords: secure multiparty computation; data privacy; epidemiological data collection; asymptotic capacity

check for updates

Citation: Cheng, J.; Liu, N.; Kang, W. On the Asymptotic Capacity of Information-Theoretic Privacy-Preserving Epidemiological Data Collection. *Entropy* **2023**, *25*, 625. https://doi.org/10.3390/e25040625

Academic Editors: Jun Chen and Sadaf Salehkalaibar

Received: 28 February 2023
Revised: 24 March 2023
Accepted: 30 March 2023
Published: 6 April 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

During any prevention and control period in an epidemic, strengthening the protection of personal information is conducive not only to safeguarding personal interests, but also better controlling the development of the epidemic. Differently from the collection of other homogeneous data, such as the large-scale labeled sample obtained in machine learning, the characteristics of medical or epidemiological data collection are reflected in the following aspects: (1) In order to establish a surveillance system for a dynamical group of people to collect syndromic data, the sample size is always changing [1,2]; (2) Data sharing and early response are critical in containing the spread of highly infectious diseases, such as COVID-19. This requires the data stored in the database to be updated to track real-time changes in symptoms, severity, or other disease-related patterns [3,4]; (3) Some of the epidemiological data, such as the locations of infected individuals, the blood oxygen saturation levels of patients with respiratory diseases after medication, or the physical condition monitoring data after viral infection, are related to the user's privacy, so a "privacy-first" approach which uses dynamic identifiers and stores their data in a cryptographically secure

manner is needed [5]. Hence, for these factors of epidemiological data collection, the key to ensuring the efficiency of information sharing and privacy protection is to maximize the balance between data privacy and collection rate in epidemic analysis.

While the data collection from public health authorities and the open-source access to researchers can provide convenience to epidemiologists, these strategies may also significantly intrude upon citizens' privacy. Some of these individuals were affected by unwanted privacy invasion, and the ubiquitous data surveillance devices certainly exacerbate those concerns. Therefore, the responsible use of shared data and algorithms, the compliance with data protection regulations, and the appropriate respect for privacy and confidentiality have become important topics in epidemiological data collection [6,7]. In 2014, the Global System for Mobile Communications (GSM) Association outlined some privacy standards for data-processing agencies regarding mobile phone data collection in response to the Ebola virus [8]. Some other methods that effectively guarantee user privacy include: encryption mechanisms [9], the privacy-aware energy-efficient framework (P-AEEF) protocol [10], the differential privacy-based classification algorithm [11], and the spatio-temporal trajectory method [12]. Moreover, unauthorized agencies, malicious hackers, and unidentified attacks, such as traffic analysis attacks, fake-node injection, and selective forwarding, may also eavesdrop on users' data under the current ambiguous and non-uniform collecting algorithms [13]. To avoid the leakage of information from malicious collecting servers or untrusted access, the privacy of data needs to be ensured during data storage and processing.

In line with ensuring the security and privacy of user data, epidemiological data collection also requires the protection of data collectors. In accessing public databases, various epidemiological investigation agencies have different requirements for data privacy. For example, many epidemiological surveys are conducted by governments, universities, research institutes, pharmaceutical companies, and private institutions. Access to public data may reveal these agencies' data preferences, resulting in information leakage. Therefore, the private-preserving epidemiological data collection includes the user's privacy regarding the storage and data collector, along with the data collector's privacy in data storage.

In epidemiological modeling, many recent studies have shown that various models have a good fitting effect on the nature of epidemics, such as the Bayesian model [14] and deep learning models, including multi-head attention, long short-term memory (LSTM), and the convolutional neural network (CNN) [15]. Additionally, when it comes to privacy and security concerns, some work conducted in computer science, cryptography, and information theory provides handy tools to model and solve such problems. The privacy leakage was modeled in differential privacy [16], k-anonymity [17], t-closeness [18], interval privacy [19], etc. With those concerns and analysis, several studies in cryptography and information theory have focused on the issue of sharing messages to untrusted agencies, such as distributed linear separable computation [20–22], secured matrix multiplication [23–25], secure aggregation [26], participatory sensing [27], and private information retrieval [28–31]. Specifically, a cryptographical epidemiological model with data security and privacy called RIPPLE was analyzed [32]. This model enables the execution of epidemiological models based on the most-recent real contact graph while keeping all the users' information private. As an extension to the model, the data collector uses the sum private information retrieval (PIR) schemes to obtain the statistical data, which is described as the sum of a set of elements from a database.

Inspired by the cryptographical epidemiological model in [32], we propose an information-theoretic privacy-preserving epidemiological-data-collection problem , described as follows. We have a large and changing number of users sharing their real-time epidemiological data to N servers. The server will update its database once new data are uploaded. The storage of those users' data is also open to all authorized data collectors. For any data collector who desires only some statistical feature, rather than all the data, it directly retrieves the statistic from N servers. We designed the protocol of the interaction between the N servers and the data collector so that when all N servers honestly answer

the queries from the data collector, then the desired statistical data can be correctly decoded at the data collector. To simplify the problem, we assume the desired statistics to be the linear combinations of users' personal data. The privacy of this model is reflected in the following three aspects: (1) The privacy of users' data against the data collector: after downloading all answers from the servers, the data collector can decode the desired data without learning anything about the irrelevant details of the users' personal information. (2) The privacy of users' data against servers: for any user successfully sharing the data to all N servers, the personal information of the user is still confidential, even if of any up to $E(E < N)$ servers collude. (3) The privacy of the data collector's preference against servers: the protocol between data collector and servers should promise that any single server cannot know any preference of the desired statistical data from the query generated by the data collector. In our paper, we take the above privacy concerns into consideration and analyze the data collector's ability of privately receiving the shared data, with respect to the number of symbols it needs to download.

The remainder of the paper is organized as follows: In Section 2, we introduce an information-theoretic description of the privacy-preserving epidemiological-data-collection problem, and we define the communication rate and capacity of our problem. In Section 3, we give a closed-form expression for the asymptotic capacity, which is the main result of the paper. In Section 4, we derive a converse proof for $E < N - 1$, which provides an upper bound on the asymptotic capacity when the number of users tends to infinity. An asymptotically capacity-achieving scheme using the technique of cross-subspace alignment when $E < N - 1$ is given in Section 5, and in Section 6, we prove that the problem is not asymptotically achievable when $E = N - 1$. Finally, in Section 7, we summarize our results and suggest some open problems in this field.

2. System Model

We formulate the 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; and the number of users ($K \in N_+$) is a large-enough integer. Here and throughout the paper, we assume that all of the random symbols in our system are generated by a large-enough finite field $\mathbb{F}_q$, and we standardize the entropy of any uniformly distributed symbol to be 1 by taking the term log in the entropy, conditional entropy, and mutual information to be base q. The model is depicted in Figure 1.

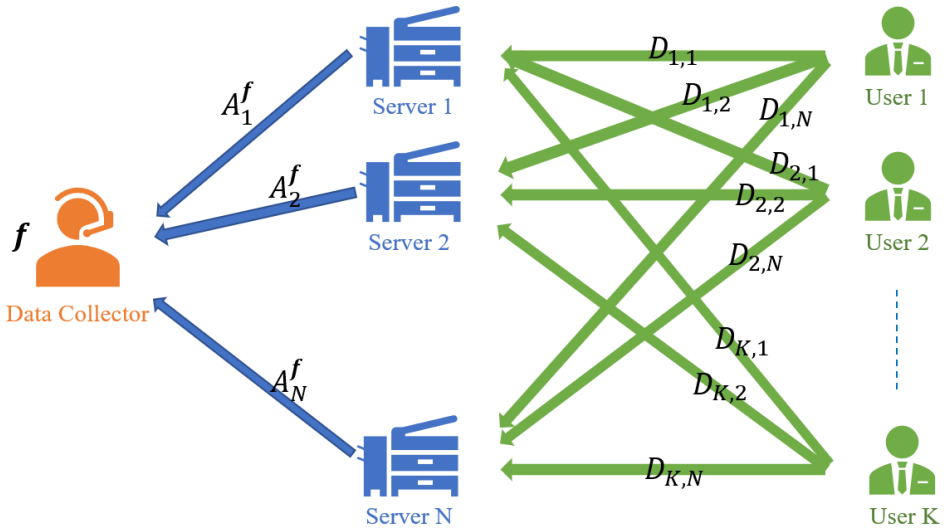

Figure 1. The secure, privacy-preserving epidemiological-data-collection problem.

Let $W_1, \cdots, W_K$ be K independent messages, where $W_k, k \in [1:K]$ denotes the epidemiological data of user k and $W_k \in GF_q^{L\times 1}$, $L \in \mathbb{N}_+$ is an $L \times 1$ vector with L i.i.d. uniform symbols from the finite field GF_q—i.e.,

$$\begin{aligned} &H(W_{[1:K]}) = \textstyle\sum_{k=1}^{K} H(W_k), \\ &H(W_k) = L, \quad \forall k \in [1:K]. \end{aligned} \tag{1}$$

The data-collection problem contains two phases: the upload phase and the computation phase. The upload phase starts when the user is required to update its data to each of the N servers. Before uploading, user k knows nothing about the contents of other users' epidemiological data. While uploading their personal data, the users would like to keep his/her message private against E ($E \leq N, E \in \mathbb{N}$) colluding servers; i.e., any of up to E servers will learn nothing about the messages uploaded by K users. For any $n \in [1:N]$, let $D_{k,n} \in \mathfrak{D}$ denote the uploaded message from the k-th user to the n-th server. We have the following equality called the privacy constraint of the users against E servers:

$$I(D_{k,\mathcal{E}}; W_k) = 0, \quad \forall k \in [1:K], \forall \mathcal{E} \subseteq [1:N], |\mathcal{E}| \leq E. \tag{2}$$

For any $k \in [1:K]$, let $Z_k \in \mathfrak{Z}_k$ denote a random variable privately generated by user k, and its realization is not known to any of the N servers. User k utilizes the user-side randomness Z_k to encipher the message W_k; i.e., for any user k, there exist N functions $\{d_{k,n}\}_{n\in[1:N]}$ such that $d_{k,n}(W_k, Z_k) = D_{k,n}$, where $d_{k,n} : GF_q^L \times \mathfrak{Z}_k \to \mathfrak{D}$ is the encoding function from the user k to the server n. We have

$$H(D_{k,[1:N]}|W_k, Z_k) = 0, \quad \forall k \in [1:K]. \tag{3}$$

When the servers receive the updated data from users, the old contents of the server will be replaced with the new contents, and all servers will be ready for the computation phase and allow the access of data updated by data collectors.

The computation phase starts when a data collector would like to compute statistical data of the current epidemiological database. To simplify our problem, we only consider statistical data to be a linear combination of all messages $W_{[K]}$. Let $W^{\mathbf{f}}$ and $\mathbf{f}$ be the statistical data and the corresponding coefficient vector. Then, we have

$$W^{\mathbf{f}} = \mathbf{f}^T \begin{bmatrix} W_1 \\ \vdots \\ W_K \end{bmatrix} = \sum_{k=1}^{K} f_k W_k, \tag{4}$$

where $W^{\mathbf{f}} \in GF_q^L$ has the same number of symbols with the message length, and $\mathbf{f} \in GF_q^K$ contains K elements in the finite field GF_q. In our setting, the coefficient vector $\mathbf{f}$ only contains the preference of the data collector among K user's epidemiological data records. Therefore, the value of $\mathbf{f}$ does not depend on the users' messages $W_{[1:K]}$ or the storage of N servers: $\{D_{k,n}\}_{k\in[1:K],n\in[1:N]}$.

Let $Z' \in \mathcal{Z}$ denote a random variable privately generated by the data collector, and its realization is not available to any of the N servers and K users. In the computation phase, the data collector with its preference vector $\mathbf{f}$ utilizes the randomness Z' to generate its queries to N servers. Assume that the query from the data collector to the n-th server is denoted as $Q_n^{\mathbf{f}} \in \mathfrak{Q}_n$. Then, the data collector uses the strategy g to map the randomness and the coefficient to the queries, such that

$$g(\mathbf{f}, Z') = \left\{Q_1^{\mathbf{f}}\ Q_2^{\mathbf{f}}\ \cdots Q_N^{\mathbf{f}}\right\}^T,$$

where $g : GF_q^K \times \mathcal{Z} \to \prod_{n=1}^N \mathfrak{Q}_n$ is the encoding function from the data collector to the servers. Hence, we have

$$H(Q_{[1:N]}^{\mathbf{f}}|Z', \mathbf{f}) = 0. \tag{5}$$

Since the randomness Z' and queries $Q_{[1:N]}^{\mathbf{f}}$ are generated privately, and the user-side data and randomness $W_{[1:K]}, Z_{[1:K]}$ are already known before the computation phase starts, the data collector has no knowledge of $W_{[1:K]}, Z_{[1:K]}$ when the queries $Q_{[1:N]}^{\mathbf{f}}$ are generated. Thus, we have

$$I(W_{[1:K]}, Z_{[1:K]}; Q_{[1:N]}^{\mathbf{f}}, Z') = 0. \tag{6}$$

Upon receiving the query $Q_n^{\mathbf{f}}$, the n-th server will send an answer $A_n^{\mathbf{f}}$ back to the data collector according to the storage of the n-th server. We assume that there is no drop-out in the model—i.e., each server n will successfully return its answer to the data collector. Let $A_n^{\mathbf{f}} \in \mathfrak{A}_n$ be the answer from the n-th server to the data collector, and then for any $n \in [1 : N]$, there exists a deterministic function $a_n^{\mathbf{f}} : \mathfrak{Q}_n \times \mathfrak{D}^K \to \mathfrak{A}_n$, such that $A_n^{\mathbf{f}} = a_n^{\mathbf{f}}(Q_n^{\mathbf{f}}, \left[D_{1,n}\, D_{2,n}, \cdots, D_{K,n}\right]^T)$, $\forall n \in [1 : N]$. Hence,

$$H(A_n^{\mathbf{f}}|Q_n^{\mathbf{f}}, D_{[1:K],n}) = 0, \quad \forall n \in [1 : N]. \tag{7}$$

We would like to design a scheme $\phi = \{d_{k,n}, g, a_n^{\mathbf{f}}\}_{k\in[1:K],n\in[1:N],\mathbf{f}\in GF_q^K}$ in the upload and computation phases so that the following three constraints can be achieved. Firstly, the data collector is able to reconstruct the desired message from all the information it has received, which we call the decodability constraint. Let ψ be the reconstruction function of the data collector, where $\psi : \prod_{n=1}^N \mathfrak{A}_n \times \prod_{n=1}^N \mathfrak{Q}_n \times GF_q^K \times \mathcal{Z} \to GF_q^L$. We have

$$\hat{W}^{\mathbf{f}} = \psi(A_{[1:N]}^{\mathbf{f}}, Q_{[1:N]}^{\mathbf{f}}, \mathbf{f}, Z'). \tag{8}$$

Let P_e be the probability of decoding error achieved by a given scheme ϕ and decoding function ψ. We obtain

$$P_e(\phi, \psi) = \max_{\mathbf{f}} \Pr\{\hat{W}^{\mathbf{f}} \neq W^{\mathbf{f}}\}.$$

According to Fano's inequality, the decodability constraint is equivalent to

$$H(W^{\mathbf{f}}|A_{[1:N]}^{\mathbf{f}}, Q_{[1:N]}^{\mathbf{f}}, Z') = o(L), \quad \forall \mathbf{f} \in GF_q^N, \tag{9}$$

when $P_e \to 0$, where $o(L)$ denotes any possible function $h : \mathbb{N}_+ \to \mathbb{R}$ of L satisfying $\lim_{L\to\infty} \frac{h(L)}{L} = 0$.

The second constraint guarantees the data privacy of K users against the data collector. The privacy leakage to the data collector is unavoidable, as the data collector must learn some statistical data of the users. This constraint requires that the data collector can learn nothing about the information of the K users other than the desired data $W^{\mathbf{f}}$. We have

$$I(W_{[1:K]}; A_{[1:N]}^{\mathbf{f}}, Q_{[1:N]}^{\mathbf{f}}, Z'|W^{\mathbf{f}}) = 0. \tag{10}$$

To protect the privacy of data collector, the third constraint requires the coefficient vector $\mathbf{f}$ of the data collector to be indistinguishable from the perspective of each server; i.e., for any different (linearly independent) coefficient vectors $\mathbf{f}$ and $\mathbf{f}'$ from the same scheme ϕ, the queries to every single server are identically distributed, so that any server cannot deduce $\mathbf{f}$ merely from the query and storage without communicating to other servers. Hence, we have

$$(A_n^{\mathbf{f}}, Q_n^{\mathbf{f}}, W_{[1:K]}) \sim (A_n^{\mathbf{f}'}, Q_n^{\mathbf{f}'}, W_{[1:K]}), \quad \forall \mathbf{f}, \mathbf{f}' \text{ linear independent} \tag{11}$$

where $A \sim B$ means that the random variables A and B have the same distribution. This constraint is called the privacy constraint of the data collector against non-colluding servers.

The reason why in the upload phase, we consider up to E servers may collude, and in the computation phase, we consider non-colluding servers to be defined by the following: 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, and the data collector may start his/her queries of certain statistics at relatively random times. Due to the dynamic topology of the servers, the numbers of colluding servers may be different during the uploading phase and the computation phase. Our work assumes that the servers are non-colluding in the computation phase, since the servers may be more interested in the epidemiological data than the data collector's interest. If $E = 1$, then we have a model where the servers are non-colluding in both the upload phase and the computation phase.

For any scheme $\phi = \{d_{k,n}, g, a_n^{\mathbf{f}}\}_{k\in[1:K],n\in[1:N],\mathbf{f}\in GF_q^K}$ that satisfies the above decodability constraint, i.e., (9), and the privacy constraints, i.e., the privacy constraint of the users against E servers (2), the privacy constraint of the users against the data collector (10), and the privacy constraint of the data collector against the non-colluding servers (11), its communication rate is characterized by the number of symbols the data collector decodes per download symbol—i.e.,

$$R := \frac{L}{\sum_{n=1}^{N} H(A_n^{\mathbf{f}})}. \tag{12}$$

It is worth noticing that R is not a function of $\mathbf{f}$ due to (11).

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

3. Main Result

For the information-theoretical framework of our privacy-preserving epidemiological-data-collection problem presented in Section 2, our main result is the characterization of the asymptotic capacity when $K \to \infty$ for any $N \in \mathbb{N}_+$ and $E \in \mathbb{N}$.

To begin with, we show that the problem is infeasible when $N = 1$ or $N = E$, since in these cases, some constraints of our setting contradict each other, and no scheme can satisfy all the constraints. Firstly, when there is only one server, the privacy of users against the data collector and the privacy of data collector against servers will conflict with each other. The reason is as follows. First, according to (11), the answer A will be given that for two different $\mathbf{f}, \mathbf{f}'$, the distribution of A is the same. Second, the decodability (9) guarantees that $W^{\mathbf{f}}$ can be derived from A. Then, $W^{\mathbf{f}'}$ can be derived from A, which contradicts the privacy of users against the data collector.

Moreover, when $E = N$, the decodability constraint and the privacy of users against the servers will conflict with each other. The reason is as follows. First, the answers $A_{[1:N]}^{\mathbf{f}}$ from N servers are given by the queries $Q_{[1:N]}^{\mathbf{f}}$ and the storage $D_{[1:K],[1:N]}$ according to (7), so there is a Markov chain $W_{[1:K]} \to (Q_{[1:N]}^{\mathbf{f}}, D_{[1:K],[1:N]}) \to A_{[1:N]}^{\mathbf{f}}$. Second, when $E = N$, $Q_{[1:N]}^{\mathbf{f}}$ and $D_{[1:K],n}$ are independent from $W_{[1:K]}$ according to (6) and (2), which contradicts the decodability constraint as $A_{[1:N]}^{\mathbf{f}}$ is independent from the database $W_{[1:K]}$. Therefore, no scheme can simultaneously satisfy all the constraints in a single-server or $E = N$ scenario, and there does not exist a positive capacity.

The following theorem gives the asymptotic capacity of private-preserving epidemiological data collection for an infinitely large K, where we have $N \geq 2$, $E < N$ servers.

Theorem 1. *Consider E as a non-negative number, and there are $N \geq 2$ servers. When $E < N$, and the number of users $K \to \infty$, the asymptotic capacity of the secure privacy-preserving epidemiological-data-collection problem is*

$$\lim_{K\to\infty, L\to\infty} C = \begin{cases} \frac{N-E-1}{N}, & \text{if } E < N-1 \\ 0, & \text{otherwise} \end{cases}. \tag{13}$$

When $E < N-1$, the converse proof of Theorem 1 will be given in Section 4, and the achievability proof will be given in Section 5 for any finite $K \in \mathbb{N}_+$. When $K \to \infty$, our scheme in Section 5 remains achievable at the same rate, and the performance of the achievability and converse will meet. When $E = N-1$, the proof of the zero asymptotic capacity is given in Section 6.

Remark 1. *From Theorem 1, we can see a threshold in the asymptotic capacity on the number of the maximized possible colluding servers E. When $E \geq N-1$, there is no scheme with a positive asymptotic capacity; when $E < N-1$, the asymptotic capacity is a decreasing function of E. When N approaches infinity while E is a constant, the asymptotic capacity approaches one.*

Remark 2. *When the number of users K is a finite integer, the achievability and converse results do not always meet. From our converse proof in Section 4, the upper bound we give for finite K depends on the value of K. However, the performance (i.e., achievable rate) of our scheme in Section 5 is irrelevant to K, even though the scheme we propose is different for the finite $K \in \mathbb{N}_+$. How to close the gap between the upper and lower bounds when K is finite is still an open problem.*

4. Proof of Theorem 1: Converse When $E < N-1$

We give the converse proof of Theorem 1 when $E < N-1$ in this section. The converse allows any feasible scheme ϕ, and we give an upper bound over the rates of all possible schemes. We start with the following lemma, which states an iterative relationship among the number of linear combinations of the users' messages.

Lemma 1. *Consider K linear independent vectors $\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_K \in GF_q^K$. Let $\mathcal{E}$ be a set and $\mathcal{E} \subseteq [1:N]$, $|\mathcal{E}| = E$. We have*

$$\begin{aligned} &H(A^{\mathbf{f}_k}_{[1:N]/\mathcal{E}} | W^{\mathbf{f}_1}, \cdots, W^{\mathbf{f}_k}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}_k}_{[1:N]}, Z') \\ &\geq \frac{L}{N-E} + \frac{1}{N-E} H(A^{\mathbf{f}_{k+1}}_{[1:N]/\mathcal{E}} | W^{\mathbf{f}_1}, \cdots, W^{\mathbf{f}_{k+1}}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}_{k+1}}_{[1:N]}, Z') - o(L). \end{aligned} \tag{14}$$

Proof. According to our problem setting, we have

$$\begin{aligned} &(N-E) H(A^{\mathbf{f}_k}_{[1:N]/\mathcal{E}} | W^{\mathbf{f}_1}, \cdots, W^{\mathbf{f}_k}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}_k}_{[1:N]}, Z') \\ \geq &\sum_{n\in[1:N]/\mathcal{E}} H(A^{\mathbf{f}_k}_n | W^{\mathbf{f}_1}, \cdots, W^{\mathbf{f}_k}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}_k}_{[1:N]}, Z') && (15) \\ = &\sum_{n\in[1:N]/\mathcal{E}} H(A^{\mathbf{f}_{k+1}}_n | W^{\mathbf{f}_1}, \cdots, W^{\mathbf{f}_k}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}_{k+1}}_{[1:N]}, Z') && (16) \\ \geq &H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}} | W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}'}_{[1:N]}, Z') \\ = &H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}} | W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}'}_{[1:N]}, Z') + H(W^{\mathbf{f}'} | A^{\mathbf{f}'}_{[1:N]/\mathcal{E}}, W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}'}_{[1:N]}, Z') - o(L) && (17) \\ = &H(W^{\mathbf{f}'} | W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}'}_{[1:N]}) + H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}} | W^{\mathbf{f}'}, W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}'}_{[1:N]}, Z') - o(L) \\ = &L + H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}} | W^{\mathbf{f}'}, W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}'}_{[1:N]}, Z') - o(L), && (18) \end{aligned}$$

where (15) holds because of the non-negativity of the following conditional entropy; i.e.,

$$H(A^{\mathbf{f}_k}_{[1:N]/(\mathcal{E}\cup\{k\})}|A^{\mathbf{f}_k}_n, W^{\mathbf{f}_1}, \cdots, W^{\mathbf{f}_k}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}_k}_{[1:N]}, Z') \geq 0, \quad \forall n \in [1:N]/\mathcal{E},$$

and (16) holds because of (11). The equality in (17) follows due to the fact that

$$H(W^{\mathbf{f}'}|A^{\mathbf{f}'}_{[1:N]/\mathcal{E}}, W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}'}_{[1:N]}, Z') = H(W^{\mathbf{f}'}|A^{\mathbf{f}'}_{[1:N]/\mathcal{E}}, W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, A^{\mathbf{f}'}_{\mathcal{E}}, Q^{\mathbf{f}'}_{[1:N]}, Z') = o(L),$$

where the first equality follows from (7), and the second equality follows from (9). Finally, (18) 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. □

The following lemma shows that any answers in a set are independent from any queries conditioned on the same set of queries to the same coefficient and any size of messages and randomnesses. This is the direct inference from the independence of message, queries, and randomnesses generated by the data collector (6).

Lemma 2. *Assume that* $\mathbf{f} \in GF_q^N$, $\mathcal{N}_1, \mathcal{N}_2 \in [1:N]$, *and* $\mathcal{K} \in [1:K]$. *Then, we have the following equality:*

$$I(A^{\mathbf{f}}_{\mathcal{N}_1}; Q^{\mathbf{f}}_{\mathcal{N}_2}|W_{\mathcal{K}}, Z', Q^{\mathbf{f}}_{\mathcal{N}_1}) = 0. \tag{19}$$

Proof. The proof is the same as Section VI, Lemma 1, in [29], and the key to this proof is that $A^{\mathbf{f}}_{\mathcal{N}_1}$ is determined by $W_{[1:K]}$, conditioned on Z' and $Q^{\mathbf{f}}_{\mathcal{N}_1}$. We omit the detailed proof here. □

The lemma below has a similar form to Lemma 2, and it shows that any set of answers with size of E do not depend on the desired statistic, conditioned on the same set of queries and the randomnesses generated by the data collector.

Lemma 3. *For any* $\mathcal{E} \subseteq [1:N], |\mathcal{E}| = E$,

$$H(A^{\mathbf{f}}_{\mathcal{E}}|Q^{\mathbf{f}}_{\mathcal{E}}, W^{\mathbf{f}}, Z') = H(A^{\mathbf{f}}_{\mathcal{E}}|Q^{\mathbf{f}}_{\mathcal{E}}, Z'). \tag{20}$$

Proof. We only need to show that $I(A^{\mathbf{f}}_{\mathcal{E}}; W^{\mathbf{f}}|Q^{\mathbf{f}}_{\mathcal{E}}, Z')$ is less than or equal to 0 because of its non-negativity.

$$\begin{aligned}
I(A^{\mathbf{f}}_{\mathcal{E}}; W^{\mathbf{f}}|Q^{\mathbf{f}}_{\mathcal{E}}, Z') \leq & I(A^{\mathbf{f}}_{\mathcal{E}}, D_{[1:K],\mathcal{E}}; W^{\mathbf{f}}|Q^{\mathbf{f}}_{\mathcal{E}}, Z') \\
= & I(D_{[1:K],\mathcal{E}}; W^{\mathbf{f}}|Q^{\mathbf{f}}_{\mathcal{E}}, Z') + I(A^{\mathbf{f}}_{\mathcal{E}}; W^{\mathbf{f}}|D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}}_{\mathcal{E}}, Z') \\
= & I(D_{[1:K],\mathcal{E}}; W^{\mathbf{f}}|Q^{\mathbf{f}}_{\mathcal{E}}, Z') \qquad (21) \\
= & H(D_{[1:K],\mathcal{E}}|Q^{\mathbf{f}}_{\mathcal{E}}, Z') - H(D_{[1:K],\mathcal{E}}|W^{\mathbf{f}}, Q^{\mathbf{f}}_{\mathcal{E}}, Z') \\
= & 0, \qquad (22)
\end{aligned}$$

where (21) holds because the answers $A^{\mathbf{f}}_{\mathcal{E}}$ are determined by $(D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}}_{\mathcal{E}}, Z')$ in (7), and (22) holds because of (6) and (2). □

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 obtain*

$$\left(1 - \frac{E}{N}\right) H(A^{\mathbf{f}}_{[1:N]}|Q^{\mathbf{f}}_{[1:N]}, Z') \geq L + H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}}_{[1:N]}, Z') - o(L). \tag{23}$$

Proof. Based on the system model, we have

$$\begin{aligned}
&H(A^{\mathbf{f}}_{[1:N]}|Q^{\mathbf{f}}_{[1:N]}, Z') \\
=&H(W^{\mathbf{f}}|Q^{\mathbf{f}}_{[1:N]}, Z') + H(A^{\mathbf{f}}_{[1:N]}|W^{\mathbf{f}}, Q^{\mathbf{f}}_{[1:N]}, Z') - H(W^{\mathbf{f}}|A^{\mathbf{f}}_{[1:N]}, Q^{\mathbf{f}}_{[1:N]}, Z') \\
=&L + H(A^{\mathbf{f}}_{[1:N]}|W^{\mathbf{f}}, Q^{\mathbf{f}}_{[1:N]}, Z') - o(L) && (24)\\
=&L + H(A^{\mathbf{f}}_{\mathcal{E}}|W^{\mathbf{f}}, Q^{\mathbf{f}}_{[1:N]}, Z') + H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, A^{\mathbf{f}}_{\mathcal{E}}, Q^{\mathbf{f}}_{[1:N]}, Z') - o(L) \\
=&L + H(A^{\mathbf{f}}_{\mathcal{E}}|W^{\mathbf{f}}, Q^{\mathbf{f}}_{\mathcal{E}}, Z') + H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, A^{\mathbf{f}}_{\mathcal{E}}, Q^{\mathbf{f}}_{[1:N]}, Z') - o(L) && (25)\\
=&L + H(A^{\mathbf{f}}_{\mathcal{E}}|Q^{\mathbf{f}}_{\mathcal{E}}, Z') + H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, A^{\mathbf{f}}_{\mathcal{E}}, Q^{\mathbf{f}}_{[1:N]}, Z') - o(L) && (26)\\
\geq&L + H(A^{\mathbf{f}}_{\mathcal{E}}|Q^{\mathbf{f}}_{\mathcal{E}}, Z') + H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}}_{[1:N]}, Z') - o(L) && (27)\\
\geq&L + \frac{E}{N}H(A^{\mathbf{f}}_{[1:N]}|Q^{\mathbf{f}}_{[1:N]}) + H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}}_{[1:N]}, Z') - o(L), && (28)
\end{aligned}$$

where (24) follows from (1), (6), and (9); (25) follows from Lemma 2 when $\mathcal{N}_1 = \mathcal{E}$ and $\mathcal{N}_2 = [1:N]$; (26) follows from (20), (6), and (7); (27) is due to (7); and (28) follows from the Han's inequality—i.e.,

$$\sum_{\mathcal{E}\subseteq[1:N],|\mathcal{E}|=E} H(A^{\mathbf{f}}_{\mathcal{E}}|Q^{\mathbf{f}}_{\mathcal{E}}) \geq \frac{E}{N}\binom{N}{E}H(A^{\mathbf{f}}_{[1:N]}|Q^{\mathbf{f}}_{[1:N]}). \tag{29}$$

□

Now, we can get the lower bound on the asymptotic download size when L and K goes infinity by applying (14) to (23). We have

$$\begin{aligned}
&\lim_{K\to\infty,L\to\infty}\left(1-\frac{E}{N}\right)H(A^{\mathbf{f}}_{[1:N]}|Q^{\mathbf{f}}_{[1:N]}) \\
\geq&\lim_{K\to\infty,L\to\infty}\left(H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}}_{[1:N]}) + L - o(L)\right) && (30)\\
=&\lim_{K\to\infty,L\to\infty} H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}}_{[1:N]}) + L && (31)\\
\geq&\lim_{K\to\infty,L\to\infty}\frac{1}{N-E}\left(L + H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, W^{\mathbf{f}'}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}'}_{[1:N]} - o(L))\right) + L \\
\geq&\left(\sum_{k=0}^{\infty}\frac{1}{(N-E)^k}\right)L, && (32)
\end{aligned}$$

where (30) follows from (23), (31) is because $o(L)$ goes zero when $L \to \infty$, and (32) follows from the non-negativity of the conditional entropy. The iteration (14) can be continued because for any $k \in \mathbb{N}$, $\frac{o(L)}{(N-E)^k} \to 0$ when $L \to \infty$. Thus, we can calculate the upper bound of the asymptotic capacity when $E < N-1$ as follows:

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

Thus, for any possible scheme ϕ satisfying the constraints of the problem, the rate cannot be more than $\frac{N-E-1}{N}$ when $E < N-1$ and $K \to \infty$.

5. Proof of Theorem 1: Achievability When $E < N - 1$

In this section, we give a cross-subspace alignment (CSA) scheme based on the coding of interference in the computation phase to reach the asymptotic capacity [33] for any integers $N > E + 1$ and $K \geq 2$. Throughout the scheme, we choose the length of each personal message $L = N - E - 1 \geq 1$, 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 $W_k^l \in GF_q$ be the l-th symbol of each W_k, $k \in [1 : K], l \in [1 : L]$, and $W^l \in GF_q^{1\times K}$ be the row vector of the l-th symbol of all K messages, i.e., $W^l = [W_1^l, \cdots, W_K^l]$. Assume that $\alpha_n, n \in [1 : N]$ are N distinct coefficients all belonging to the set $\{\alpha \in GF_q : \alpha + i \neq 0, i \in [1 : L]\}$—i.e., for any $i, j \in [1 : N], \alpha_i \neq \alpha_j$. Note that the α_ns are globally shared variables known to the users, servers, and the data collector. In order to protect the privacy of the users against the servers, each user k will generate $L \times E$ random noises Z_{le}^k uniformly from GF_q. The uploaded information to the n-th server by the k-th user is given by

$$D_{k,n} = d_{k,n}(W_k, Z_k) = \begin{bmatrix} W_k^1 + \sum_{e=1}^{E}(1+\alpha_n)^e Z_{1e}^k \\ \vdots \\ W_k^L + \sum_{e=1}^{E}(L+\alpha_n)^e Z_{Le}^k \end{bmatrix}^T, \quad \forall k \in [1 : K], n \in [1 : N], \tag{33}$$

For convenience of notation, we write the content stored at server n in a vector form as

$$D_n = [D_{1,n}^1, \cdots, D_{K,n}^1, D_{1,n}^2, \cdots, D_{K,n}^2, \cdots, D_{1,n}^L, \cdots, D_{K,n}^L] \tag{34}$$

$$= \begin{bmatrix} W^1 + \sum_{e=1}^{E}(1+\alpha_n)^e Z_{1e} \\ \vdots \\ W^L + \sum_{e=1}^{E}(L+\alpha_n)^e Z_{Le}. \end{bmatrix}^T, \tag{35}$$

where $D_n \in GF_q^{1\times KL}$, and Z_{le} is defined as $Z_{le} = [Z_{le}^1, \cdots, Z_{le}^K]$.

In the computation phase, the query to Server n is determined by the coefficient $\mathbf{f}$ and the randomness from the data collector Z'. We design the query to Server n based on $\mathbf{f}$ as

$$Q_n^{\mathbf{f}} = \begin{bmatrix} \frac{\Delta_n}{1+\alpha_n}(\mathbf{f} + (1+\alpha_n)Z_1') \\ \vdots \\ \frac{\Delta_n}{L+\alpha_n}(\mathbf{f} + (L+\alpha_n)Z_L') \end{bmatrix}, \tag{36}$$

where $Z_1', \cdots, Z_L'$ are L random column vectors of length K, whose elements are uniformly distributed on GF_q, generated by the data collector.

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

$$A_n^{\mathbf{f}} = a_n^{\mathbf{f}}(Q_n^{\mathbf{f}}, D_{[1:K],n}) = D_n \cdot Q_n^{\mathbf{f}} \tag{37}$$

$$= \left(W^1 + \sum_{e=1}^{E}(1+\alpha_n)^e Z_{1e}\right) \cdot \left(\frac{\Delta_n}{1+\alpha_n}(\mathbf{f} + (1+\alpha_n)Z_1')\right) + \cdots + \left(W^L + \sum_{e=1}^{E}(L+\alpha_n)^e Z_{Le}\right) \cdot \left(\frac{\Delta_n}{L+\alpha_n}(\mathbf{f} + (L+\alpha_n)Z_L')\right), \forall n \in [1 : N]. \tag{38}$$

According to the expansion of the representation (38) in the descending power of α, we can see that $\frac{A_n^{\mathbf{f}}}{\Delta_n}$ is the sum of $\sum_{l=1}^{L}\frac{1}{l+\alpha_n}W^l \cdot \mathbf{f}$ and a polynomial of degree E in α_n. We can rewrite (38) as

$$A_n^{\mathbf{f}} = \Delta_n \left(\sum_{l=1}^{L} \frac{W^l \mathbf{f}}{l + \alpha_n} + \sum_{e=0}^{E} I_e \alpha_n^e \right), \tag{39}$$

where I_e is the coefficient of α_n^e for any $e \in [0 : E]$. We can clearly see from (38) and (39) that I_e is not a function of n. If we write the answers to the data collector from all the N servers together, we can get the following formula in a matrix form:

$$\begin{bmatrix} \frac{A_1^{\mathbf{f}}}{\Delta_1} \\ \frac{A_2^{\mathbf{f}}}{\Delta_2} \\ \cdots \\ \frac{A_N^{\mathbf{f}}}{\Delta_N} \end{bmatrix} = \begin{bmatrix} \frac{1}{1+\alpha_1} & \cdots & \frac{1}{L+\alpha_1} & 1 & \alpha_1 & \cdots & \alpha_1^E \\ \frac{1}{1+\alpha_2} & \cdots & \frac{1}{L+\alpha_2} & 1 & \alpha_2 & \cdots & \alpha_2^E \\ & & & \cdots & & & \\ \frac{1}{1+\alpha_N} & \cdots & \frac{1}{L+\alpha_N} & 1 & \alpha_N & \cdots & \alpha_N^E \end{bmatrix} \cdot \begin{bmatrix} W^1 \cdot \mathbf{f} \\ \vdots \\ W^L \cdot \mathbf{f} \\ I_0 \\ \vdots \\ I_E \end{bmatrix}. \tag{40}$$

Now, we prove that this scheme satisfies the decodability constraint, i.e., (9), and the privacy constraints—i.e., the privacy constraint of the users against E servers (2), the privacy constraint of the users against the data collector (10), and the privacy constraint of the data collector against the non-colluding servers (11).

Recall that in our scheme, we let $L = N - E - 1$. The decodability constraint is satisfied because the matrix in (40) is an $N \times N$ full-rank matrix when the α_ns are distinct Lemma 5 in [33]. Hence, $W^1 \cdot \mathbf{f}, \cdots, W^L \cdot \mathbf{f}$ can be fully recovered from $\frac{A_1^{\mathbf{f}}}{\Delta_1}, \cdots, \frac{A_N^{\mathbf{f}}}{\Delta_N}$, and the desired data $W^{\mathbf{f}}$ in (4) is obtained by

$$W^{\mathbf{f}} = W^T \cdot \mathbf{f} = \begin{bmatrix} W^1 \cdot \mathbf{f} \\ \cdots \\ W^L \cdot \mathbf{f} \end{bmatrix}. \tag{41}$$

The privacy constraint of the users against E colluding servers is satisfied due to the sharing strategy of the users. In (33), we know that the k-th user shares its l-th symbol to the n-th server in a form

$$D_{k,n}^l = W_k^l + \sum_{e=1}^{E} (1 + \alpha_n)^e Z_{le}^k, \tag{42}$$

where $D_{k,n}^l$ denotes the storage in server n that W_k^l shares. The security needs to guarantee that any E out of N servers do not know W_k for any $k \in [1 : K]$. For any set $\mathcal{E} := \{o_1, \cdots, o_E\}$ such that $\mathcal{E} \in [1 : N]$ and $|\mathcal{E}| = E$, we choose the E servers to be in $\mathcal{E}$. We can write the storage of these servers with respect to what W_k^l shares in a matrix form:

$$\begin{bmatrix} D_{k,o_1}^l \\ \cdots \\ D_{k,o_E}^l \end{bmatrix} = \begin{bmatrix} W_k^l \\ \cdots \\ W_k^l \end{bmatrix} + \begin{bmatrix} l + \alpha_{o_1} & (l + \alpha_{o_1})^2 & \cdots & (l + \alpha_{o_1})^E \\ & & \cdots & \\ l + \alpha_{o_E} & (l + \alpha_{o_E})^2 & \cdots & (l + \alpha_{o_E})^E \end{bmatrix} \cdot \begin{bmatrix} Z_{l1}^k \\ \cdots \\ Z_{lE}^k \end{bmatrix}. \tag{43}$$

Notice that the Vandermonde matrix in (43), denoted by $V_{\mathcal{E}}$, is invertible for distinct $\{1 + \alpha_e : \alpha_e \in GF_q, e \in \mathcal{E}\}$, so the second term of (43) contains E symbols that are linearly independent. The privacy constraint of the users against E colluding servers can be guaranteed as the E additional random symbols protect the shared message. We have

$$
\begin{aligned}
& I(D_{k,\mathcal{E}}; W_k) \\
\leq & \sum_{i=1}^{L} \sum_{j=1}^{L} I(D_{k,\mathcal{E}}^{i}; W_k^j | D_{k,\mathcal{E}}^{[1:i-1]}; W_k^{[1:j-1]}) \\
= & \sum_{i=1}^{L} \sum_{j=1}^{L} I(W_k^i \mathbf{1}_E + V_E \cdot Z_{i,\mathcal{E}}^k; W^j | D_{[1:i-1],\mathcal{E}}; W^{[1:j-1]}) \qquad (44) \\
= & \sum_{l=1}^{L} I\left(W_k^l; W_k^l \mathbf{1}_E + V_E \cdot Z_{l,\mathcal{E}}^k\right) \qquad (45) \\
= & \sum_{l=1}^{L} I\left(W_k^l; Z_{l,\mathcal{E}}^k\right) \\
= & 0, \qquad \forall k \in [1:K], n \in [1:N].
\end{aligned}
$$

where (44) comes from (43), and (45) holds because when $i < j$, we have

$$
\begin{aligned}
& I(W_k^i \mathbf{1}_E + V_E \cdot Z_{i,\mathcal{E}}^k; W^j | D_{[1:i-1],\mathcal{E}}; W^{[1:j-1]}) \\
= & H(W_k^i \mathbf{1}_E + V_E \cdot Z_{i,\mathcal{E}}^k | D_{[1:i-1],\mathcal{E}}; W^{[1:j-1]}) - H(W_k^i \mathbf{1}_E + V_E \cdot Z_{i,\mathcal{E}}^k | D_{[1:i-1],\mathcal{E}}; W^{[1:j]}) \\
= & H(W_k^i \mathbf{1}_E + V_E \cdot Z_{i,\mathcal{E}}^k | W^i) - H(W_k^i \mathbf{1}_E + V_E \cdot Z_{i,\mathcal{E}}^k | W^i) \\
= & 0, \quad \forall i \in [1:L], \forall j \in [i+1, L],
\end{aligned}
$$

and when $i > j$, we have

$$
\begin{aligned}
& I(W_k^i \mathbf{1}_E + V_E \cdot Z_{i,\mathcal{E}}^k; W^j | D_{[1:i-1],\mathcal{E}}; W^{[1:j-1]}) \\
= & H(W_k^i \mathbf{1}_E + V_E \cdot Z_{i,\mathcal{E}}^k | D_{[1:i-1],\mathcal{E}}; W^{[1:j-1]}) - H(W_k^i \mathbf{1}_E + V_E \cdot Z_{i,\mathcal{E}}^k | D_{[1:i-1],\mathcal{E}}; W^{[1:j]}) \\
= & H(W_k^i \mathbf{1}_E + V_E \cdot Z_{i,\mathcal{E}}^k) - H(W_k^i \mathbf{1}_E + V_E \cdot Z_{i,\mathcal{E}}^k) \\
= & 0, \quad \forall i \in [1:L], \forall j \in [1:i-1],
\end{aligned}
$$

so the remaining items are those (i, j) such that $i = j = l$.

To prove that the privacy constraint of the data collector against the non-colluding servers is satisfied, we notice that the query to each server is composed of the desired coefficient $\mathbf{f}$ and independent additional randomness $Z'_{[1:L]}$. Consider two linear independent vectors, $\mathbf{f}, \mathbf{f}' \in GF_q^K$. For any $l \in [1:L]$, the l-th entries of $Q_n^{\mathbf{f}}$ and $Q_n^{\mathbf{f}'}$ are $\frac{\Delta_n}{l+\alpha_n}(\mathbf{f} + (l+\alpha_n)Z'_l)$ and $\frac{\Delta_n}{l+\alpha_n}(\mathbf{f}' + (l+\alpha_n)Z'_l)$, respectively. Notice that Z'_l (thus $\frac{\Delta_n}{l+\alpha_n} \cdot Z'_l$) is chosen uniformly from GF_q^K, and that $\frac{\Delta_n}{l+\alpha_n} \cdot \mathbf{f}$ and $\frac{\Delta_n}{l+\alpha_n} \cdot \mathbf{f}'$ are two deterministic vectors in GF_q^K. The l-th symbol of $Q_n^{\mathbf{f}}$ or $Q_n^{\mathbf{f}'}$ has the same distribution. Therefore, any single server can not distinguish queries from the data collector with one coefficient $\mathbf{f}$. Furthermore, if we assume the desired coefficient $\mathbf{t}$ is a random variable with some distribution known only to the data collector, and $\mathbf{f}$ is an implementation of $\mathbf{t}$, we can prove that the mutual information between $\mathbf{t}$ and $(Q_n^{\mathbf{t}}, A_n^{\mathbf{t}}, W_{[1:K]})$ is zero. We have

$$
\begin{aligned}
& I(Q_n^{\mathbf{t}}, A_n^{\mathbf{t}}, W_{[1:K]}; \mathbf{t}) \\
\leq & I(Q_n^{\mathbf{t}}, W_{[1:K]}, \{Z_{le}\}_{l\in[1:L],e\in[1:E]}; \mathbf{t}) && (46)\\
= & I(Q_n^{\mathbf{t}}; \mathbf{t} | W_{[1:K]}, \{Z_{le}\}_{l\in[1:L],e\in[1:E]}) && (47)\\
= & H(Q_n^{\mathbf{t}} | W_{[1:K]}, \{Z_{le}\}_{l\in[1:L],e\in[1:E]}) - H(Q_n^{\mathbf{t}} | \mathbf{t}, W_{[1:K]}, \{Z_{le}\}_{l\in[1:L],e\in[1:E]}) \\
= & H(Q_n^{\mathbf{t}}) - H\left(\begin{bmatrix} \mathbf{t} + (1+\alpha_n)(Z_1') \\ \cdots \\ \mathbf{t} + (L+\alpha_n)(Z_L') \end{bmatrix} \middle| \mathbf{t}, W_{[1:K]}, \{Z_{le}\}_{l\in[1:L],e\in[1:E]} \right) && (48)\\
= & H(Q_n^{\mathbf{t}}) - H(Z_1', \cdots, Z_L') \\
= & L - L \\
= & 0,
\end{aligned}
$$

where (46) is from (38), (47) is from (10), and (48) is from (6).

Finally, to prove that the privacy constraint of the users against the data collector is satisfied, we construct a basis of GF_q^K containing the desired $\mathbf{f}$. Assume that 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

$$
\begin{aligned}
& I\left(W_{[1:K]}; A_{[1:N]}^{\mathbf{f}}, Q_{[1:N]}^{\mathbf{f}}, Z' | W^{\mathbf{f}}\right) \\
= & \sum_{l=1}^{L} I\left(W^l; A_{[1:N]}^{\mathbf{f}}, Q_{[1:N]}^{\mathbf{f}}, Z' | W^{[1:l-1]}, W^{\mathbf{f}}\right) \\
\leq & \sum_{l=1}^{L} I\left(W^l; A_{[1:N]}^{\mathbf{f}}, Q_{[1:N]}^{\mathbf{f}}, Z' | W^{[1:L]/\{l\}}, Z, W^{\mathbf{f}}\right) && (49)\\
= & \sum_{l=1}^{L} I\left(\left(W^l + \sum_{e=1}^{E} (l+\alpha_n)^e Z_{le} \right) \cdot \right. \\
& \qquad \left. \left(\frac{\Delta_n}{1+\alpha_n} (\mathbf{f} + (l+\alpha_n)(Z_l')^T) \right)_{n\in[1:N]}; W^l | W^{[1:L]/\{l\}}, Z, W^{\mathbf{f}} \right) && (50)\\
= & \sum_{l=1}^{L} I\left(\left(W^l (Z_l')^T + \sum_{e=1}^{E} (l+\alpha_n)^e Z_{le} (Z_l')^T \right)_{n\in[1:N]}; W^l | W^{[1:L]/\{l\}}, W^{\mathbf{f}} \right) && (51)\\
\leq & \sum_{l=1}^{L} I\left(\left(W^l (Z_l')^T, Z_{le} (Z_l')^T \right)_{e\in[1:E]}; W^l | W^{[1:L]/\{l\}}, W^{\mathbf{f}} \right) && (52)\\
= & 0, && (53)
\end{aligned}
$$

where (49) holds because $(W^{[1:L]/\{l\}}, Z)$ is independent with W^l; (50) holds because except for the term containing W^l, all terms in (38) are given, so deducting them would not change the mutual information; (51) is because Δ_n is a constant; (52) is because for any $n \in [1:N]$ and $e \in [1:E]$, $(l+\alpha_n)^e$ is a constant, and for $n \in [1:N]$, $W^l (Z_l')^T + \sum_{e=1}^{E} (l+\alpha_n)^e Z_{le} (Z_l')^T$ can be derived by $W^l (Z_l')^T$ and $\left(Z_{le} (Z_l')^T\right)_{e\in[1:E]}$; and (53) holds because for any $l \in [1:L]$, $\{Z_{le}\}_{e\in[1:E]}$ and Z_l' are generated independently of $W^{[1:L]}$.

Thus, we prove that the scheme satisfies all the constraints. As the answer from any server contains one symbol (the inner product) from GF_q, the rate of our proposed scheme is

$$
R = \frac{L}{\sum_{n=1}^{N} H(A_n^{\mathbf{f}})} \frac{N-E-1}{N}. \tag{54}
$$

It can be seen that the scheme we construct by (33), (36), and (38) is achievable with the rate (54) invariant with K by rearranging the data $W_{[1:K]}$ to $W^{[1:L]}$. 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 $K \to \infty$.

6. Converse Result When $E = N - 1$

In this section, we prove that the asymptotic capacity is zero when $N = E + 1$. First, the inequality (23) also holds when $N = E + 1$ because the inequality in (15) becomes an equality. Hence, we have

$$H(A^{\mathbf{f}}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, D_{[1:K],\mathcal{E}}, S, Q^{\mathbf{f}}_{[1:N]}, Z') \geq L + H(A^{\mathbf{f}'}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}}, W^{\mathbf{f}'}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}'}_{[1:N]}, Z') - o(L). \tag{55}$$

Thus, for any linear independent vectors $\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_K \in GF_q^K$, we have

$$\begin{aligned}
\left(1 - \frac{E}{N}\right) H(A^{\mathbf{f}}_{[1:N]}|Q^{\mathbf{f}}_{[1:N]}) \geq & H(A^{\mathbf{f}_1}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}_1}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}_1}_{[1:N]}, Z') + L - o(L) && (56)\\
\geq & 2L + H(A^{\mathbf{f}_2}_{[1:N]/\mathcal{E}}|W^{\mathbf{f}_1}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}_2}_{[1:N]}, Z') - o(L) && (57)\\
\geq & \sum_{k=0}^{K} L + H(A^{\mathbf{f}_K}_{[1:N]/\mathcal{E}}|W^{[1:L]}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}_K}_{[1:N]}, Z') && (58)\\
= & \sum_{k=0}^{K} L + H(A^{\mathbf{f}_K}_{[1:N]/\mathcal{E}}|W_{[1:K]}, D_{[1:K],\mathcal{E}}, Q^{\mathbf{f}_K}_{[1:N]}, Z') && (59)\\
= & KL,
\end{aligned}$$

where (56), (57), and (58) are from (55); and (59) holds because any K linear independent vectors constitute a basis in GF_q^K, so $W_{[1:K]}$ can be decoded. We can see that the download will go to infinity when $K \to \infty$, and thus the asymptotic capacity will be

$$\begin{aligned}
\lim_{K\to\infty, L\to\infty} C \leq & \lim_{K\to\infty, L\to\infty} \frac{L}{H(A^{\mathbf{f}}_{[1:N]})}\\
\leq & \lim_{K\to\infty, L\to\infty} \frac{L}{H(A^{\mathbf{f}}_{[1:N]}|H(Q^{\mathbf{f}}_{[1:N]})}\\
\leq & \lim_{K\to\infty, L\to\infty} \frac{1 - \frac{E}{N}}{KL}\\
= & 0.
\end{aligned}$$

The upper bound of C indicates that it is unfeasible to construct a scheme that has a positive communication rate when $K \to \infty$. However, differently from the $N = E$ scenario, our proof of the zero asymptotic capacity does not mean that there does not exist any scheme that satisfies all the constraints of our problem. In other words, there may be schemes that can satisfy all the constraints of the problem, albeit with an asymptotic capacity of zero. The detailed construction of a feasible scheme for $E = N - 1$ and finite K is still an open problem.

7. Conclusions

Thanks to the research on data privacy and modeling of infectious diseases, the privacy-preserving epidemiological-data-collection problem was proposed, which aims to maximize the collection rate while protecting the privacy of all users' data and the data collector's preferences. We have found the asymptotic capacity of this problem, and the result shows that when there is more than one remaining server that not colluding with the other servers to decode the users' data, the asymptotic capacity exists. The objective of this work was to find the best performance for the privacy-preserving epidemiological-data-collection problem, and we partly achieved this goal by giving the construction and proof of the optimal scheme when K is infinitely large. The achievability for $N \geq E + 2$ was given by the cross-subspace alignment method, and the infeasibility of $N = 1$ or $N = E$ was also proved. Our findings not only extend the research on secure multi-party computing systems in information theory, but also provide information-theoretic frame-

works, implementations, and capacity bounds for the study of privacy epidemiological modeling. Although we characterized the asymptotic capacity for this problem, the exact capacity is still unknown. Some future directions include finding the exact capacity for finite K, the construction of a scheme when $N = E + 1$, and the performance of asymptotic capacity under irregular colluding patterns. In general, our result of asymptotic capacity for this problem will provide useful insights for further studies in data both privacy and epidemiological modeling.

Author Contributions: Conceptualization, J.C., N.L. and W.K.; methodology, J.C., N.L. and W.K.; formal analysis, J.C., N.L. and W.K.; writing—original draft preparation, J.C.; writing—review and editing, J.C., N.L. and W.K.; supervision, N.L. and W.K.; Funding acquisition, N.L. and W.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the National Natural Science Foundation of China under Grants 61971135 and 62071115, and the Research Fund of National Mobile Communications Research Laboratory, Southeast University (No. 2023A03).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

COVID-19	Corona Virus Disease 2019
GSM	Global System for Mobile Communications
P-AEEF	Privacy-Aware Energy-Efficient Framework
LSTM	Long Short-Term Memory
CNN	Convolutional Neural Network
CSA	Cross Subspace Alignment

References

1. Kim, J.; Kwon, O. A Model for Rapid Selection and COVID-19 Prediction with Dynamic and Imbalanced Data. *Sustainability* **2021**, *13*, 3099. [CrossRef]
2. Olson, D.; Lamb, M.; Lopez, M.R.; Colborn, K.; Paniagua-Avila, A.; Zacarias, A.; Zambrano-Perilla, R.; Rodríguez-Castro, S.R.; Cordon-Rosales, C.; Asturias, E.J. Performance of a Mobile Phone App-Based Participatory Syndromic Surveillance System for Acute Febrile Illness and Acute Gastroenteritis in Rural Guatemala. *J. Med. Internet Res.* **2017**, *19*, e368. [CrossRef] [PubMed]
3. Demirci, J.R.; Bogen, D.L. An Ecological Momentary Assessment of Primiparous Women's Breastfeeding Behavior and Problems From Birth to 8 Weeks. *J. Hum. Lact.* **2017**, *33*, 285–295.
4. Silva de Lima, A.L.; Hahn, T.; Evers, L.J.W.; de Vries, N.M.; Cohen, E.; Afek, M.; Bataille, L.; Daeschler, M.; Claes, K.; Boroojerdi, B.; et al. Feasibility of large-scale deployment of multiple wearable sensors in Parkinson's disease. *PLoS ONE* **2017**, *12*, e0189161. [CrossRef] [PubMed]
5. Fahey, R.A.; Hino, A. COVID-19, digital privacy, and the social limits on data-focused public health responses. *Int. J. Inf. Manag.* **2020**, *55*, 102181. [CrossRef]
6. Ienca, M.; Vayena, E. On the responsible use of digital data to tackle the COVID-19 pandemic. *Nat. Med.* **2020**, *26*, 463–464. [CrossRef]
7. Marabelli, M.; Vaast, E.; Li, J.L. Preventing the digital scars of COVID-19. *Eur. J. Inf. Syst.* **2021**, *30*, 176–192. [CrossRef]
8. GSM Association. GSMA Guidelines on the Protection of Privacy in the Use of Mobile Phone Data for Responding to the Ebola Outbreak. 2014. Available online: https://www.gsma.com/mobilefordevelopment/resources/gsma-guidelines-on-the-protection-of-privacy-in-the-use-of-mobile-phone-data-for-responding-to-the-ebola-outbreak/ (accessed on 15 January 2023).
9. Pahlevanzadeh, B.; Koleini, S.; Fadilah, S.I. Security in IoT: Threats and Vulnerabilities, Layered Architecture, Encryption Mechanisms, Challenges and Solutions. In Proceedings of the Advances in Cyber Security, Penang, Malaysia, 24–25 August 2021; Anbar, M., Abdullah, N., Manickam, S., Eds.; Springer: Singapore, 2021; pp. 267–283.
10. Al-Turjman, F.; Deebak, B. Privacy-Aware Energy-Efficient Framework Using the Internet of Medical Things for COVID-19. *IEEE Internet Things Mag.* **2020**, *3*, 64–68. [CrossRef]

11. Fan, W.; He, J.; Guo, M.; Li, P.; Han, Z.; Wang, R. Privacy preserving classification on local differential privacy in data centers. *J. Parallel Distrib. Comput.* **2020**, *135*, 70–82. [CrossRef]
12. Su, J.; He, X.; Qing, L.; Niu, T.; Cheng, Y.; Peng, Y. A novel social distancing analysis in urban public space: A new online spatio-temporal trajectory approach. *Sustain. Cities Soc.* **2021**, *68*, 102765. [CrossRef]
13. Muhammad, M.H.G.; Alyas, T.; Ahmad, F.; Butt, F.H.; Qazi, W.M.; Saqib, S. An analysis of security challenges and their perspective solutions for cloud computing and IoT. *EAI Endorsed Trans. Scalable Inf. Syst.* **2020**, *8*. [CrossRef]
14. Anderson, S.C.; Edwards, A.M.; Yerlanov, M.; Mulberry, N.; Stockdale, J.E.; Iyaniwura, S.A.; Falcão, R.C.; Otterstatter, M.C.; Irvine, M.A.; Janjua, N.Z.; et al. Quantifying the impact of COVID-19 control measures using a Bayesian model of physical distancing. *PLoS Comput. Biol.* **2020**, *16*, e1008274. [CrossRef]
15. Abbasimehr, H.; Paki, R. Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization. *Chaos Solitons Fractals* **2021**, *142*, 110511. [CrossRef]
16. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In Proceedings of the Theory of Cryptography, New York, NY, USA, 4–7 March 2006; Halevi, S., Rabin, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 265–284.
17. Sweeney, L. k-anonymity: A model for protecting privacy. *Int. J. Uncertainty Fuzziness-Knowl.-Based Syst.* **2002**, *10*, 557–570. [CrossRef]
18. Li, N.; Li, T.; Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 15–20 April 2007; pp. 106–115. [CrossRef]
19. Ding, J.; Ding, B. Interval Privacy: A Framework for Privacy-Preserving Data Collection. *IEEE Trans. Signal Process.* **2022**, *70*, 2443–2459. [CrossRef]
20. Wan, K.; Sun, H.; Ji, M.; Caire, G. Distributed Linearly Separable Computation. *IEEE Trans. Inf. Theory* **2021**, *68*, 1259–1278. [CrossRef]
21. Wan, K.; Sun, H.; Ji, M.; Caire, G. On the Tradeoff Between Computation and Communication Costs for Distributed Linearly Separable Computation. *IEEE Trans. Commun.* **2021**, *69*, 7390–7405. [CrossRef]
22. Wan, K.; Sun, H.; Ji, M.; Caire, G. On Secure Distributed Linearly Separable Computation. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 912–926. [CrossRef]
23. Chen, Z.; Jia, Z.; Wang, Z.; Jafar, S.A. GCSA Codes with Noise Alignment for Secure Coded Multi-Party Batch Matrix Multiplication. *IEEE J. Sel. Areas Inf. Theory* **2021**, *2*, 306–316. [CrossRef]
24. Chang, W.T.; Tandon, R. On the Capacity of Secure Distributed Matrix Multiplication. In Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 9–13 December 2018; pp. 1–6. [CrossRef]
25. Hasırcıoğlu, B.; Gómez-Vilardebó, J.; Gündüz, D. Bivariate Polynomial Codes for Secure Distributed Matrix Multiplication. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 955–967. [CrossRef]
26. Zhao, Y.; Sun, H. Information Theoretic Secure Aggregation with User Dropouts. In Proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT), Melbourne, Australia, 12–20 July 2021; pp. 1124–1129. [CrossRef]
27. Chen, Q.; Zheng, S.; Weng, Z. Data Collection with Privacy Preserving in Participatory Sensing. In Proceedings of the 2017 IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS), Shenzhen, China, 15–17 December 2017; pp. 49–56. [CrossRef]
28. Sun, H.; Jafar, S.A. The Capacity of Private Information Retrieval. *IEEE Trans. Inf. Theory* **2017**, *63*, 4075–4088. [CrossRef]
29. Wang, Q.; Sun, H.; Skoglund, M. The capacity of private information retrieval with eavesdroppers. *IEEE Trans. Inf. Theory* **2018**, *65*, 3198–3214. [CrossRef]
30. Yao, X.; Liu, N.; Kang, W. The Capacity of Private Information Retrieval Under Arbitrary Collusion Patterns for Replicated Databases. *IEEE Trans. Inf. Theory* **2021**, *67*, 6841–6855. [CrossRef]
31. Cheng, J.; Liu, N.; Kang, W.; Li, Y. The Capacity of Symmetric Private Information Retrieval Under Arbitrary Collusion and Eavesdropping Patterns. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 3037–3050. [CrossRef]
32. Günther, D.; Holz, M.; Judkewitz, B.; Möllering, H.; Pinkas, B.; Schneider, T. *PEM: Privacy-Preserving Epidemiological Modeling*; Report 2020/1546; The International Association for Cryptologic Research Location: Hyderabad, India, 2020; p. 1546.
33. Jia, Z.; Sun, H.; Jafar, S.A. Cross Subspace Alignment and the Asymptotic Capacity of X-Secure T-Private Information Retrieval. *IEEE Trans. Inf. Theory* **2019**, *65*, 5783–5798. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.