Private Aggregation from Fewer Anonymous Messages

Badih Ghazi[1], Pasin Manurangsi[1], Rasmus Pagh[1,2], and Ameya Velingker[1]

[1] Google Research, Mountain View CA 94043, USA
[2] IT University of Copenhagen, Denmark
{badihghazi,pasin,pagh,ameyav}@google.com

Abstract. Consider the setup where n parties are each given an element x_i in the finite field $\mathbb{F}_q$ and the goal is to compute the sum $\sum_i x_i$ in a secure fashion and with as little communication as possible. We study this problem in the *anonymized model* of Ishai et al. (FOCS 2006) where each party may broadcast anonymous messages on an insecure channel.

We present a new analysis of the one-round "split and mix" protocol of Ishai et al. In order to achieve the same security parameter, our analysis reduces the required number of messages by a $\Theta(\log n)$ multiplicative factor.

We also prove lower bounds showing that the dependence of the number of messages on the domain size, the number of parties, and the security parameter is essentially tight.

Using a reduction of Balle et al. (2019), our improved analysis of the protocol of Ishai et al. yields, in the same model, an (ε, δ)-differentially private protocol for aggregation that, for any constant $\varepsilon > 0$ and any $\delta = \frac{1}{\mathrm{poly}(n)}$, incurs only a constant error and requires only a *constant number of messages* per party. Previously, such a protocol was known only for $\Omega(\log n)$ messages per party.

Keywords: Secure Aggregation · Anonymous Channel · Shuffled Model · Differential Privacy.

1 Introduction

We study one-round multi-party protocols for the problem of secure aggregation: Each of n parties holds an element of the field $\mathbb{F}_q$ and we wish to compute the sum of these numbers, while satisfying the security property that for every two inputs with the same sum, their transcripts are "indistinguishable." The protocols we consider work in the *anonymized model*, where parties are able to send anonymous messages through an insecure channel and indistinguishability is in terms of the *statistical distance* between the two transcripts (i.e., this is information-theoretic security rather than computational security). This model was introduced by Ishai et al. [17] in their work on cryptography from

anonymity[3]. We refer to [17, 8] for a discussion of cryptographic realizations of an anonymous channel.

The secure aggregation problem in the anonymized model was studied already by Ishai et al. [17], who gave a very elegant one-round "split and mix" protocol. Under their protocol, each party i holds a private input x_i and sends m anonymized messages consisting of random elements of $\mathbb{F}_q$ that are conditioned on summing to x_i. Upon receiving these mn anonymized messages from n parties, the server adds them up and outputs the result. Pseudocode of this protocol is shown as Algorithm 1. Ishai et al. [17] show that as long as m exceeds a threshold of $\Theta(\log n + \sigma + \log q)$, this protocol is σ-*secure* in the sense that the statistical distance between transcripts resulting from inputs with the same sum is at most $2^{-\sigma}$.

Differentially Private Aggregation in the Shuffled Model. An exciting recent development in differential privacy is the *shuffled model*, which is closely related to the aforementioned anononymized model. The shuffled model provides a middle ground between two widely-studied models of differential privacy. In the *central model*, the data structure released by the analyst is required to be differentially private, whereas the *local model* enforces the more stringent requirement that the messages sent by each party be private. While protocols in the central model generally allow better accuracy, they require a much greater level of trust to be placed in the analyzer, an assumption that may be unsuitable for certain applications. The *shuffled model* is based on the Encode-Shuffle-Analyze architecture of [6] and was first analytically studied by [12, 8] and further studied in recent work [4, 13]. It seeks to bridge the two aforementioned models and assumes the presence of a trusted shuffler that randomly permutes all incoming messages from the parties before passing them to the analyzer (see Section 2 for formal definitions.) The shuffled model is particularly compelling because it allows the possibility of obtaining more accurate communication-efficient protocols than in the local model while placing far less trust in the analyzer than in the central model. Indeed, the power of the shuffled model has been illustrated by a number of recent works that have designed algorithms in this model for a wide range of problems such as privacy amplification, histograms, heavy hitters, and range queries [8, 12, 4, 13, 11].

The appeal of the shuffled model provides the basis for our study of differentially private protocols for aggregation in this work. Most relevant to the present work are the recent differentially private protocols for aggregation of real numbers in the shuffled model provided by [8, 4, 15, 3]. The strongest of these results [3] shows that an extension of the split and mix protocol yields an (ε, δ)-differentially private protocol for aggregation with error $O(1 + 1/\varepsilon)$ and $m = O(\log(n/\delta))$ messages, each consisting of $O(\log n)$ bits.

[3] Ishai et al. in fact considered a more general model in which the adversary is allowed to corrupt some of the parties; please refer to the discussion at the end of Section 1.1 for more details.

1.1 Our Results

Upper bound. We prove that the split and mix protocol is in fact secure for a much smaller number of messages. In particular, for the same security parameter σ, the number of messages required in our analysis is $\Theta(\log n)$ times smaller than the bound in [17]:

Theorem 1 (Improved upper bound for split and mix). *Let n and q be positive integers and σ be a positive real number. The split and mix protocol (Algorithm 1 and [17]) with n parties and inputs in $\mathbb{F}_q$ is σ-secure for m messages, where $m = O\left(1 + \frac{\sigma + \log q}{\log n}\right)$.*

An interesting case to keep in mind is when the field size q and the inverse statistical distance 2^σ are bounded by a polynomial in n. In this case, Theorem 1 implies that the protocol works already with a *constant* number of messages, improving upon the known $O(\log n)$ bound.

Lower bound. We show that, in terms of the number of messages m sent by each party, Theorem 1 is essentially tight not only for just the split and mix protocol but also for *every* one-round protocol.

Theorem 2 (Lower bound for every one-round protocol). *Let n and q be positive integers, and $\sigma \geq 1$ be a real number. In any σ-secure, one-round aggregation protocol over $\mathbb{F}_q$ in the anonymized model, each of the n parties must send $\Omega\left(1 + \frac{\sigma}{\log(\sigma n)} + \frac{\log q}{\log n}\right)$ messages.*

The lower bound holds regardless of the message size and asymptotically matches the upper bound under the very mild assumption that σ is bounded by a polynomial in n. Furthermore, when σ is larger, the bound is tight up to a factor $O\left(\frac{\log \sigma}{\log n}\right)$.

We point out that Theorem 2 provides a nearly-tight lower bound on the *number of messages*. In terms of the total communication per party, improvements are still possible when $\sigma + \log q = \omega(\log n)$. We discuss this further, along with other interesting open questions, in Section 5.

Corollary for Differentially Private Aggregation. As stated earlier, the differentially private aggregation protocols of [3, 15] both use extensions of the split and mix protocol. Moreover, Balle et al. use the security guarantee of the split and mix protocol as a blackbox and derive a differential privacy guarantee from it [3, Lemma 4.1]. Specifically, when ε is a constant and $\delta \geq \frac{1}{\mathrm{poly}(n)}$, their proof uses the split and mix protocol with field size $q = \mathrm{poly}(n)$. Previous analyses required $m = \Omega(\log n)$; however, our analysis works with a *constant* number of messages. In general, Theorem 1 implies (ε, δ)-differential privacy with a factor $\Theta(\log n)$ fewer messages than known before:

Corollary 1 (Differentially private aggregation in the shuffled model). *Let n be a positive integer, and let ε, δ be positive real numbers. There is an (ε, δ)-differentially private aggregation protocol in the shuffled model for inputs in $[0,1]$ having absolute error $O(1+1/\varepsilon)$ in expectation, using $O\left(1+\frac{\log(1/\delta)}{\log n}\right)$ messages per party, each consisting of $O(\log n)$ bits.*

A more comprehensive comparison between our differentially private aggregation protocol in Corollary 1 and previous protocols is presented in Figure 1.

We end this subsection by remarking that Ishai et al. [17] in fact considered a setting that is more general than what we have described so far. Specifically, they allow the adversary to corrupt a certain number of parties. In addition to the transcript of the protocol, the adversary knows the input and messages of these corrupted parties. (Alternatively, one can think of these corrupted parties as if they are colluding to learn the information about the remaining parties.) As already observed in [17], the security of the split and mix protocol still holds in this setting except that n is now the number of honest (i.e., uncorrupted) parties. In other words, Theorem 1 remains true in this more general setup but with n being the number of honest parties instead of the total number of parties.

Discussion and comparison of parallel and subsequent work. Concurrently and independently of our work, Balle et al. [2, 5] obtained an upper bound that is asymptotically the same as the one in Theorem 1. They also give explicit constants, whereas we state our theorem in asymptotic notation and do not attempt to optimize the constants in our proof.

A key difference between our work and theirs is that in addition to the analysis of the split and mix protocol, we manage to prove a matching lower bound on the required number of messages for any protocol (see Theorem 2), which establishes the near-tightness of the algorithmic guarantees in our upper bound. Our lower bound approach could potentially be applied to other problems pertaining to the anonymous model and possibly differential privacy.

The upper bound proofs use different techniques. Balle et al. reduce the question to an analysis of the number of connected components of a certain random graph, while our proof analyzes the rank deficiency of a carefully-constructed random matrix. While the upper bound of Balle et al. is shown for summation over any abelian group, our proofs are presented for finite fields. We note, though, that our lower bound proof carries over verbatim to any abelian group.

A subsequent work [14] obtained an $(\varepsilon, 0)$-differentially private aggregation protocol with error $O_\varepsilon(1)$ and where each user sends $O_\varepsilon(\log^3 n)$ messages each consisting of $O(\log\log n)$ bits (see Figure 1 for the explicit bounds).

1.2 Applications and Related Work

At first glance it may seem that aggregation is a rather limited primitive for combining data from many sources in order to analyze it. However, in important approaches to machine learning and distributed/parallel data processing,

the mechanism for combining computations of different parties is *aggregation of vectors*. Since we can build vector aggregation in a straightforward way from scalar aggregation, our results can be applied in these settings.

Before discussing this in more detail, we mention that it is shown in [17] that summation protocols can be used as building blocks for realizing *general* secure computations in a specific setup where a server mediates computation of a function on data held by n other parties. However, the result assumes a somewhat weak security model (see in Appendix D of [17] for more details).

Machine Learning and Data Analytics. Secure aggregation has applications in so-called *federated* machine learning [21] (see, e.g., [18] for a recent survey). The idea is to train a machine learning model without collecting data from any party, and instead compute weight updates in a distributed manner by sending model parameters to all parties, locally running stochastic gradient descent on private data, and aggregating model updates over all parties. For learning algorithms based on gradient descent, a secure aggregation primitive can be used to compute global weight updates without compromising privacy [23, 24]. It is known that gradient descent can work well even if data is accessible only in noised form, in order to achieve differential privacy (e.g., [1, 25]).

Beyond gradient descent, as observed in [8], we can translate any *statistical query* over a distributed data set to an aggregation problem over numbers in $[0, 1]$. That is, every learning problem solvable using a small number of statistical queries [19] can be solved privately and efficiently based on secure aggregation.

Moreover, very recent work in eye-tracking research [22, 29] study differential privacy for eye-tracking tasks, the most basic of which is the *aggregation* of users' gaze maps.

Sketching. Research in the area of data stream algorithms has uncovered many non-trivial algorithms that are compact *linear sketches*, see, e.g., [9, 31]. As noted already in [17], linear sketches can be implemented using secure aggregation by computing linear sketches locally, and then using aggregation to compute their sum which yields the sketch of the whole dataset. Typically, linear sketches do not reveal much information about their input, and are robust to the noise needed to ensure differential privacy, though specific guarantees depend on the sketch in question. We refer to [20, 26, 27] for examples and further discussion.

Secure aggregation protocols. Secure aggregation protocols are well-studied, both under cryptographic assumptions and with respect to differential privacy. We refer to the survey of Goryczka et al. [16] for an overview, but note that our approach leads to protocols that use less communication than existing (multi-round) protocols. The trust assumptions needed for implementing a shuffler (e.g., using a mixnet) are, however, slightly different from the assumptions typically used for secure aggregation protocols. Practical secure aggregation typically relies on an honest-but-curious assumption, see e.g. [7]. In that setting, such protocols typically require five rounds of communication with $\Omega(n)$ bits of communication and $\Omega(n^2)$ computation per party. A more recent work [28] using

Reference	#messages / n	Message size	Expected error
Cheu et al. [8]	$\varepsilon\sqrt{n}$ ℓ	1	$\frac{1}{\varepsilon}\log\frac{n}{\delta}$ $\sqrt{n}/\ell + \frac{1}{\varepsilon}\log\frac{1}{\delta}$
Balle et al. [4]	1	$\log n$	$\frac{n^{1/6}\log^{1/3}(1/\delta)}{\varepsilon^{2/3}}$
Ghazi et al. [15]	$\log(\frac{n}{\varepsilon\delta})$	$\log(\frac{n}{\delta})$	$\frac{1}{\varepsilon}\sqrt{\log\frac{1}{\delta}}$
Balle et al. [3]	$\log(\frac{n}{\delta})$	$\log n$	$\frac{1}{\varepsilon}$
This work (Corollary 1)	$1 + \frac{\log(1/\delta)}{\log n}$	$\log n$	$\frac{1}{\varepsilon}$
Ghazi et al. [14] ($\delta = 0$)	$\frac{\log^3 n}{\varepsilon}$	$\log\log n$	$\frac{\sqrt{\log(1/\varepsilon)}}{\varepsilon^{3/2}}$

Fig. 1. Comparison of differentially private aggregation protocols in the shuffled model with (ε, δ)-differential privacy. The number of parties is n, and ℓ is an integer parameter. Message sizes are in bits. For readability, we assume that $\varepsilon \leq O(1)$, and asymptotic notations are suppressed.

homomorphic threshold encryption gives a protocol with three messages and constant communication and computation per party in addition to a (reusable) two-message setup (consisting of $\Omega(n)$ communication per party). By contrast, our aggregation protocol has a single round of constant communication and computation per party, albeit in the presence of a trusted shuffler. We note that for an apples to apples comparison of these approaches, one would need to look at actual implementations of the shuffler which is beyond the scope of this work.

Other related models. A very recent work [30] has designed an extension of the shuffled model, called *Multi Uniform Random Shufflers* and analyzed its trust model and privacy-utility tradeoffs. Since they consider a more general model, our differentially private aggregation protocol would hold in their setup as well.

There has also been work on aggregation protocols in the multiple servers setting, e.g., the PRIO system [10]; here the protocol is secure as long as at least one server is honest. Thus trust assumptions of PRIO are somewhat different from those underlying shuffling and mixnets. While each party would be able to check the output of a shuffler, to see if its message is present, such a check is not possible in the PRIO protocol making server manipulation invisible even if the number of parties is known. On the other hand, PRIO handles malicious parties that try to manipulate the result of a summation by submitting illegal data — a challenge that has not been addressed yet for summation in the shuffled model but that would be interesting future work.

1.3 The Split and Mix Protocol

The protocol of [17] is shown in Algorithm 1. To describe the main guarantee proved in [17] regarding Algorithm 1, we need some notation. For any input sequence $\mathbf{x} \in \mathbb{F}_q^n$, we denote by $\mathcal{S}_\mathbf{x}$ the distribution on $\mathbb{F}_q^{mn}$ obtained by sampling

$y_{m(i-1)+1}, \ldots, y_{mi} \in \mathbb{F}_q$ uniformly at random conditioned on $y_{m(i-1)+1} + \cdots + y_{mi} = x_i$, sampling a random permutation $\pi : [mn] \to [mn]$, and outputting $(y_{\pi(1)}, \ldots, y_{\pi(mn)})$. Ishai et al. [17] proved that for some $m = O(\log n + \sigma + \log q)$ and for any two input sequences $\mathbf{x}, \mathbf{x}' \in \mathbb{F}_q^n$ having the same sum (in $\mathbb{F}_q$), the distributions $\mathcal{S}_{\mathbf{x}}$ and $\mathcal{S}_{\mathbf{x}'}$ are $2^{-\sigma}$-close in statistical distance.

Algorithm 1: Split and mix encoder from [17]

Input: $x \in \mathbb{F}_q$, positive integer parameter m
Output: Multiset $\{y_1, \ldots, y_m\} \subseteq \mathbb{F}_q$

for $j = 1, \ldots, m-1$ **do**
  $y_j \leftarrow \text{Uniform}(\mathbb{F}_q)$
$y_m \leftarrow x - \sum_{j=1}^{m-1} y_j$ (in $\mathbb{F}_q$)
return $\{y_1, \ldots, y_m\}$

1.4 Overview of Proofs

We now give a short overview of the proofs of Theorems 1 and 2. For ease of notation, we define $\mathcal{B}_s$ to be the set of all input vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{F}_q^n$ with a fixed sum $x_1 + x_2 + \cdots + x_n = s$.

Upper Bound. To describe the main idea behind our upper bound, we start with the following notation. For every $x \in \mathbb{F}_q$, we denote by $\mathcal{S}_x$ the uniform distribution on $\mathbb{F}_q^{mn}$ conditioned on all coordinates summing to x.

To prove Theorem 1, we have to show that for any two input sequences $\mathbf{x}, \mathbf{x}' \in \mathbb{F}_q^n$ such that $\sum_{i\in[n]} x_i = \sum_{i\in[n]} x'_i$, the statistical distance between $\mathcal{S}_{\mathbf{x}}$ and $\mathcal{S}_{\mathbf{x}'}$ is at most $\gamma = 2^{-\sigma}$. By the triangle inequality, it suffices to show that the statistical distance between $\mathcal{S}_{\mathbf{x}}$ and $\mathcal{S}_{x_1+\cdots+x_n}$ is at most $\gamma/2$. (Theorem 3). Note that $\mathcal{S}_{x_1+\cdots+x_n}$ puts equal mass on all vectors in $\mathbb{F}_q^{mn}$ whose sum is equal to $x_1 + \cdots + x_n$. Thus, our task boils down to showing that the mass put by $\mathcal{S}_{\mathbf{x}}$ on a random sample from $\mathcal{S}_{x_1+\cdots+x_n}$ is well-concentrated. We prove this via a second order method (specifically, Chebyshev's inequality). This amounts to computing the mean and bounding the variance. The former is a simple calculation whereas the latter is more technically involved and reduces to proving a probabilistic bound (Theorem 4) on the rank deficit of a certain random matrix (specified in Definitions 7 and 8). A main ingredient in the proof of this bound is a combinatorial characterization (Lemma 2) of the rank deficit of the relevant matrices in terms of *matching partitions* (given in Definition 9).

Lower Bound. For the lower bound (Theorem 2), our proof consists of two parts: a "security-dependent" lower bound $m \geq \Omega\left(\frac{\sigma}{\log(\sigma n)}\right)$ and a "field-dependent" lower bound $m \geq \Omega\left(\frac{\log q}{\log n}\right)$. Combining these two yields Theorem 2. We start

by outlining the field-dependent bound as it is simpler before we outline the security-dependent lower bound which is technically more challenging.

Field-Dependent Lower Bound. To prove the field-dependent lower bound (formally stated in Theorem 5), the key idea is to show that for any $s \in \mathbb{F}_q$, there exist distinct inputs $\mathbf{x}, \mathbf{x}' \in \mathcal{B}_s$ such that the statistical distance between $\mathcal{S}_{\mathbf{x}}$ and $\mathcal{S}_{\mathbf{x}'}$ is at least $1 - n^{nm}/q^{n-1}$ (see Lemma 4). We do so by proving the same quantitative lower bound on the *average* statistical distance between $\mathcal{S}_{\mathbf{x}}$ and $\mathcal{S}_{\mathbf{x}'}$ over all pairs $\mathbf{x}, \mathbf{x}' \in \mathcal{B}_s$.

The average statistical distance described above can be written as the sum, over all $\mathbf{y}$, of the average difference in probability mass assigned to $\mathbf{y}$ by $\mathbf{x}$ and $\mathbf{x}'$. Thus, we consider how to lower bound this coordinate-wise probability mass difference for an arbitrary $\mathbf{y}$.

There are at most n^{nm} ways to associate each of the nm elements of $\mathbf{y}$ with a particular party. Since any individual party's encoding uniquely determines the corresponding input, it follows that any shuffled output $\mathbf{y}$ could have arisen from at most n^{nm} inputs $\mathbf{x}$. Moreover, since there are exactly q^{n-1} input vectors $\mathbf{x} \in \mathcal{B}_s$, it follows that there are at least $q^{n-1} - n^{nm}$ possible inputs $\mathbf{x} \in \mathcal{B}_s$ that cannot possibly result in $\mathbf{y}$ as an output. This implies that the average coordinate-wise probability mass difference, over all $\mathbf{x}, \mathbf{x}' \in \mathcal{B}_s$, is at least $\left(1 - \frac{q^{n-1}}{n^{nm}}\right)$ times the average probability mass assigned to $\mathbf{y}$ over all inputs in $\mathcal{B}_s$. Summing this up over all $\mathbf{y}$ yields the desired bound.

Security-Dependent Lower Bound. To prove the security-dependent lower bound, it suffices to prove the following statement (see Theorem 7): if Enc is the encoder of any aggregation protocol in the anonymized model for $n > 2$ parties with m messages sent per party, then there is a vector $\mathbf{x} \in \mathcal{B}_0$ such that the statistical distance between the distributions of the shuffled output $\mathbf{y}$ corresponding to inputs $\mathbf{0}$ and $\mathbf{x}$ is at least $\frac{1}{(10nm)^{5m}}$.

Let us first sketch a proof for the particular case of the split and mix protocol. In this case, we set $\mathbf{x} = (\underbrace{1, 1, \ldots, 1}_{n-1}, -(n-1))$, and we will bound from below the statistical distance by considering the "distinguisher" $\mathcal{A}$ which chooses a random permutation $\pi : [nm] \to [nm]$ and accepts iff $y_{\pi(1)} + \cdots + y_{\pi(m)} = 0$. We can argue (see Subsection 4.2) that the probability that $\mathcal{A}$ accepts under the distribution $\mathcal{S}_{\mathbf{0}}$ is larger by an additive factor of $\frac{1}{(en)^m}$ than the probability that it accepts under the distribution $\mathcal{S}_{\mathbf{x}}$. To generalize this idea to arbitrary encoders (beyond Ishai et al.'s protocol), it is natural to consider a distinguisher which accepts iff $y_{\pi(1)}, \ldots, y_{\pi(m)}$ is a valid output of the encoder when the input is zero. Unlike the case of Ishai et al., in general when $\pi(1), \ldots, \pi(m)$ do not all come from the same party, it is not necessarily true that the acceptance probability would be the same for both distributions. To circumvent this, we pick the smallest integer t such that the t-message marginal of the encoding of 0 and that of input 1 are substantially different, and we let the distinguisher perform an analogous check on $y_{\pi(1)}, \ldots, y_{\pi(t)}$ (instead of $y_{\pi(1)}, \ldots, y_{\pi(m)}$ as before). Another complication

that we have to deal with is that we can no longer consider the input vector $(1, \cdots, 1, -(n-1))$ as in the lower bound for Ishai et al.'s protocol sketched above. This is because the t-message marginal of the encoding of $-(n-1)$ could deviate from that for input 0 more substantially than from that for input 1, which could significantly affect the acceptance probability. Hence, to overcome this issue, we instead set x^* to the minimizer of this value t among all elements of $\mathbb{F}_q$, and use the input vector $\mathbf{x} = (x^*, \ldots, x^*, -(n-1)x^*)$ (for more details we refer the reader to the full proof in Subsection 4.2).

Organization of the Rest of the Paper

We start with some preliminaries in Section 2. We prove our main upper bound (Theorem 1) in Section 3. We prove our lower bound (Theorem 2) in Section 4. The proof of Corollary 1 appears in Appendix B.

2 Preliminaries

2.1 Protocols

In this paper, we are concerned with answering the question of how many messages are needed for protocols to achieve certain security or cryptographic guarantees. We formally define the notion of protocols in the models of interest to us.

We first define the notion of a *secure protocol* in the *shuffled model*. An n-user *secure protocol* in the *shuffled model*, $\mathcal{P} = (\mathsf{Enc}, \mathcal{A})$, consists of a randomized *encoder* (also known as *local randomizer*) $\mathsf{Enc} : \mathcal{X} \to \mathcal{Y}^m$ and an *analyzer* $\mathcal{A} : \mathcal{Y}^{nm} \to \mathcal{Z}$. Here, $\mathcal{Y}$ is known as the *message alphabet*, $\mathcal{Y}^m$ is the *message space* for each user, and $\mathcal{Z}$ is the *output space* of the protocol. The protocol $\mathcal{P}$ implements the following mechanism: each party i holds an input $x_i \in \mathcal{X}$ and encodes x_i as Enc_{x_i}. (Note that Enc_{x_i} is possibly random based on the private randomness of party i.) The concatenation of the encodings, $\mathbf{y} = (\mathsf{Enc}_{x_1}, \mathsf{Enc}_{x_2}, \ldots, \mathsf{Enc}_{x_n}) \in \mathcal{Y}^{nm}$ is then passed to a trusted *shuffler*, who chooses a uniformly random permutation π on nm elements and applies π to $\mathbf{y}$. The output is submitted to the analyzer, which then outputs $\mathcal{P}(\mathbf{x}) = \mathcal{A}(\pi(\mathbf{y})) \in \mathcal{Z}$.

In this paper, we will be concerned with protocols for *aggregation*, in which $\mathcal{X} = \mathcal{Z} = \mathbb{F}_q$ (a finite field on q elements) and $\mathcal{Y} = [\ell] = \{1, 2, \ldots, \ell\}$, and

$$\mathcal{A}(\pi(\mathsf{Enc}_{x_1}, \mathsf{Enc}_{x_2}, \ldots, \mathsf{Enc}_{x_n})) = \sum_{i=1}^{n} x_i,$$

i.e., the protocol always outputs the sum of the parties' inputs, regardless of the randomness over the encoder and the shuffler.

A related notion that we consider in this work is a one-round protocol $\mathcal{P} = (\mathsf{Enc}, \mathcal{A})$ in the *anonymized model*. The notion is similar to that of a secure protocol in the shuffled model except that there is no shuffler. Rather, the

analyzer $\mathcal{A}$ receives a *multiset* of nm messages obtained by enumerating all m messages of each of the n parties' encodings. It is straightforward to see that the two models are equivalent, in the sense that a protocol in one model works in the other and the distributions of the view of the analyzer are the same.

2.2 Distributions Related to a Protocol

To study a protocol and determine its security and privacy, it is convenient to define notations for several probability distributions related to the protocol. First, we use $\mathcal{E}_x^{\mathsf{Enc}}$ to denote the distribution of the (random) encoding of x:

Definition 1. *For a protocol $\mathcal{P}$ with encoding function Enc, we let $\mathcal{E}_x^{\mathsf{Enc}}$ denote the distribution of outputs over $\mathcal{Y}^m$ obtained by applying Enc to $x \in \mathcal{X}$.*

Furthermore, for a vector $\mathbf{x} \in \mathcal{X}^n$, we use $\mathcal{E}_{\mathbf{x}}^{\mathsf{Enc}}$ to denote the distribution of the concatenation of encodings of $x_1, \ldots, x_n$, as stated more formally below.

Definition 2. *For an n-party protocol $\mathcal{P}$ with encoding function Enc and $\mathbf{x} \in \mathcal{X}^n$, we let $\mathcal{E}_{\mathbf{x}}^{\mathsf{Enc}}$ denote the distribution over $\mathcal{Y}^{nm}$ obtained by applying Enc individually to each element of $\mathbf{x}$, i.e.,*

$$\mathcal{E}_{\mathbf{x}}^{\mathsf{Enc}} \sim \left(\mathcal{E}_{x_1}^{\mathsf{Enc}}, \mathcal{E}_{x_2}^{\mathsf{Enc}}, \ldots, \mathcal{E}_{x_n}^{\mathsf{Enc}}\right).$$

Finally, we define $\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}$ to be $\mathcal{E}_{\mathbf{x}}^{\mathsf{Enc}}$ after random shuffling. Notice that $\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}$ is the distribution of the transcript seen at the analyzer.

Definition 3. *For an n-party protocol $\mathcal{P}$ with encoding function Enc and $\mathbf{x} \in \mathcal{X}^n$, we let $\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}$ denote the distribution over $\mathcal{Y}^{nm}$ obtained by applying Enc to the elements of $\mathbf{x}$ and then shuffling the resulting nm-tuple, i.e.,*

$$\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}} \sim \pi \circ \mathcal{E}_{\mathbf{x}}^{\mathsf{Enc}}$$

for π a uniformly random permuation over nm elements.

2.3 Security and Privacy

Given two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$, we let $\mathrm{SD}(\mathcal{D}_1, \mathcal{D}_2)$ denote the *statistical distance* (aka the total variation distance) between $\mathcal{D}_1$ and $\mathcal{D}_2$.

We begin with a notion of σ-security for computation of a function f, which essentially says that distinct inputs with a common function value should be (almost) indistinguishable:

Definition 4 (σ-security). *An n-user one-round protocol $\mathcal{P} = (\mathsf{Enc}, \mathcal{A})$ in the anonymized model is said to be σ-secure for computing a function $f : \mathcal{X}^n \to \mathcal{Z}$ if for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$ such that $f(\mathbf{x}) = f(\mathbf{x}')$, we have*

$$\mathrm{SD}\left(\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}, \mathcal{S}_{\mathbf{x}'}^{\mathsf{Enc}}\right) \leq 2^{-\sigma}.$$

In this paper, we will primarily be concerned with the function that sums the inputs of each party, i.e., $f : \mathbb{F}_q^n \to \mathbb{F}_q$ given by $f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^n x_i$.

We now define the notion of (ε, δ)-*differential privacy*. We say that two input vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ and $\mathbf{x}' = (x'_1, x'_2, \ldots, x'_n) \in \mathcal{X}^n$ are *neighboring* if they differ on at most one party's data, i.e., $x_i = x'_i$ for all but one value of i.

Definition 5 ((ε, δ)-differential privacy). *An algorithm $M : \mathcal{X}^* \to \mathcal{Z}$ is (ε, δ)-differentially private if for every neighboring input vectors $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$ and every $S \subseteq \mathcal{Z}$, we have*

$$\Pr[M(\mathbf{x}) \in S] \leq e^{\varepsilon} \cdot \Pr[M(\mathbf{x}') \in S] + \delta,$$

where probability is over the randomness of M.

We now define (ε, δ)-differential privacy specifically in the *shuffled model*.

Definition 6. *A protocol $\mathcal{P}$ with encoder $\mathsf{Enc} : \mathcal{X} \to \mathcal{Z}^m$ is (ε, δ)-differentially private in the shuffled model if the algorithm $M : \mathcal{X}^n \to \mathcal{Z}^{nm}$ given by*

$$M(x_1, x_2, \ldots, x_n) = \pi(\mathsf{Enc}_{x_1}, \mathsf{Enc}_{x_2}, \ldots, \mathsf{Enc}_{x_n})$$

is (ε, δ)-differentially private, where π is a uniformly random permutation on nm elements.

3 Proof of Theorem 1

In this section, we prove Theorem 1, i.e., that the split and mix protocol of Ishai et al. is σ-secure even for $m = \Theta\left(1 + \frac{\sigma + \log q}{\log n}\right)$ messages, improving upon the known bounds of $O(\log n + \sigma + \log q)$ [17, 3, 15].

Since we only consider Ishai et al.'s split and mix protocol in this section, we will drop the superscript from $\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}$ and simply write $\mathcal{S}_{\mathbf{x}}$ to refer to the shuffled output distribution of the protocol. Recall that, by the definition of the protocol, $\mathcal{S}_{\mathbf{x}}$ is generated as follows: for every $i \in [n]$, sample $y_{m(i-1)+1}, \ldots, y_{mi} \in \mathbb{F}_q$ uniformly at random conditioned on $y_{m(i-1)+1} + \cdots + y_{mi} = x_i$. Then, pick a random permutation $\pi : [mn] \to [mn]$ and output $(y_{\pi(1)}, \ldots, y_{\pi(mn)})$.

Showing that the protocol is σ-secure is by definition equivalent to showing that $\mathrm{SD}(\mathcal{S}_{\mathbf{x}}, \mathcal{S}_{\mathbf{x}'}) \leq 2^{-\sigma}$ for all inputs $\mathbf{x}, \mathbf{x}' \in \mathbb{F}_q^n$ such that $\sum_{i \in [n]} x_i = \sum_{i \in [n]} x_i$.

In fact, we prove a stronger statement, that each $\mathcal{S}_{\mathbf{x}}$ is γ-close (in statistical distance) to the distribution that is uniform over all vectors in $\mathbb{F}_q^{mn}$ whose sum of all coordinates is equal to $\sum_{i \in [n]} x_i$, as stated below.

Theorem 3. *For every $a \in \mathbb{F}_q$, let $\mathcal{S}_a$ denote the distribution on $\mathbb{F}_q^{mn}$ generated uniformly at random conditioned on all coordinates summing to a. For any parameter $\gamma > 0$ and any $m \geq \Theta(1 + \log_n(q/\gamma))$, the following holds: for every $\mathbf{x} \in \mathbb{F}_q^n$, the statistical distance between $\mathcal{S}_{\mathbf{x}}$ and $\mathcal{S}_{x_1 + \cdots + x_n}$ is at most γ.*

When plugging in $\gamma = 2^{-\sigma-1}$, Theorem 3 immediately implies Theorem 1 via the triangle inequality.

We now outline the overall proof approach. First, observe that $\mathcal{S}_{x_1+\cdots+x_n}$ puts probability mass equally across all vectors $\mathbf{t} \in \mathbb{F}_q^{mn}$ whose sum of all coordinates is $x_1+\cdots+x_n$, whereas $\mathcal{S}_{\mathbf{x}}$ puts mass proportional to the number of permutations $\pi : [mn] \to [mn]$ such that $\mathbf{y} := (t_{\pi^{-1}(1)}, \ldots, t_{\pi^{-1}(mn)})$ satisfies $y_{m(i-1)+1} + \cdots + y_{mi} = x_i$ for all $i \in [n]$. Thus, our task boils down to proving that this latter number is well-concentrated (for a random $\mathbf{t} \in \mathrm{supp}(\mathcal{S}_{x_1+\cdots+x_n})$). We prove this via a second moment method (specifically Chebyshev's inequality). Carrying this out amounts to computing the first moment and upper-bounding the second moment of this number. The former is a simple calculation, whereas the latter involves proving an inequality regarding the rank of a certain random matrix (Theorem 4). We do so by providing a combinatorial characterization of the rank deficit of the relevant matrices (Lemma 2).

The rest of this section is organized as follows. In Subsection 3.1, we define appropriate random variables, state the bound we want for the second moment (Lemma 4), and show how it implies our main theorem (Theorem 3). Then, in Subsection 3.2, we relate the second moment to the rank of a random matrix (Proposition 1). Finally, we give a probabilistic bound on the rank of such a random matrix in Subsection 3.3 (Theorem 4).

3.1 Bounding Statistical Distance via Second Moment Method

From now on, let us fix $\mathbf{x} \in \mathbb{F}_q^n$, and let $a = x_1 + \cdots + x_n$. The variables we define below will depend on $\mathbf{x}$ (or a), but, for notational convenience, we avoid indicating these dependencies in the variables' names.

For every $\mathbf{t} \in \mathbb{F}_q^{mn}$, let $Z_{\mathbf{t}}$ denote the number of permutations $\pi : [mn] \to [mn]$ such that $t_{\pi(m(i-1)+1)} + \cdots + t_{\pi(mi)} = x_i$ for all $i \in [n]$. From the definition[4] of $\mathcal{S}_{\mathbf{x}}$, its probability mass function is

$$f_{\mathcal{S}_{\mathbf{x}}}(\mathbf{t}) = \frac{Z_{\mathbf{t}}}{(mn)! \cdot q^{(m-1)n}}. \tag{1}$$

As stated earlier, Theorem 3 is essentially about the concentration of $Z_{\mathbf{t}}$, which we will prove via the second moment method. To facilitate the proof, for every $\pi : [mn] \to [mn]$, let us also denote by $Y_{\mathbf{t},\pi}$ the indicator variable of "$t_{\pi(r(i-1)+1)} + \cdots + t_{\pi(ri)} = x_i$ for all $i \in [n]$". Note that by definition we have

$$Z_{\mathbf{t}} = \sum_{\pi \in \Pi_{mn}} Y_{\mathbf{t},\pi} \tag{2}$$

where Π_{mn} denotes the set of all permutations of $[mn]$.

When we think of $\mathbf{t}$ as a random variable distributed according to $\mathcal{S}_a$, the mean of $Y_{\mathbf{t},\pi}$ (and hence of $Z_{\mathbf{t}}$) can be easily computed: the probability that $\mathbf{t}$

[4] Note that, if derived directly from the definition of $\mathcal{S}_{\mathbf{x}}$, π here should be replaced by π^{-1}. However, these two definitions are equivalent since $\pi \mapsto \pi^{-1}$ is a bijection.

satisfies "$t_{\pi(m(i-1)+1)} + \cdots + t_{\pi(mi)} = x_i$" is exactly $1/q$ for each $i \in [n-1]$, and these events are independent. Furthermore, when these events are true, it is automatically the case that the condition holds for $i = n$. Hence, we immediately have:

Observation 1. *For every $\pi \in \Pi_{mn}$,*

$$\mathop{\mathbb{E}}_{\mathbf{t}\sim\mathcal{S}_a}[Y_{\mathbf{t},\pi}] = \frac{1}{q^{n-1}}. \tag{3}$$

The more challenging part is upper-bounding the second moment of $Z_{\mathbf{t}}$ (where we once again think of $\mathbf{t}$ as a random variable drawn from $\mathcal{S}_a$). This is equivalent to upper-bounding the expectation of $Y_{\mathbf{t},\pi} \cdot Y_{\mathbf{t},\pi'}$, where π, π' are independent uniformly random permutations of $[mn]$ and $\mathbf{t}$ is once again drawn from $\mathcal{S}_a$. On this front, we will show the following bound in the next subsections.

Lemma 1. *For every $\pi \in \Pi_{mn}$, we have*

$$\mathop{\mathbb{E}}_{\pi,\pi'\sim\Pi_{mn},\mathbf{t}\sim\mathcal{S}_a}[Y_{\mathbf{t},\pi} \cdot Y_{\mathbf{t},\pi'}] \leq \sum_{k\geq 1} \frac{q^k}{q^{2n-1}} \cdot \left(\frac{n^2}{(n/2)^{m-2}}\right)^{\frac{k-1}{2}}. \tag{4}$$

Since there are many parameters, the bound might look a bit confusing. However, the only property we need in order to show concentration of $Z_{\mathbf{t}}$ is that the right-hand side of (4) is dominated by the $k = 1$ term. This is the case when the term inside the parenthesis is $q^{-\Omega(1)}$, which indeed occurs when $m \geq 4 + \Omega(\log_n q)$.

The bound in Lemma 1 will be proved in the subsequent sections. For now, let us argue why such a bound implies our main theorem (Theorem 3).

Proof of Theorem 3. First, notice that (2) and Observation 1 together imply that

$$\mathop{\mathbb{E}}_{\mathbf{t}\sim\mathcal{S}_a}[Z_{\mathbf{t}}] = \frac{(mn)!}{q^{n-1}}. \tag{5}$$

For convenience, let us define μ as $\frac{(mn)!}{q^{n-1}}$.

We now bound the second moment of $Z_{\mathbf{t}}$ as follows:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{t}\sim\mathcal{S}_a}[Z_{\mathbf{t}}^2] &= \mathop{\mathbb{E}}_{\mathbf{t}\sim\mathcal{S}_a}\left[\left(\sum_{\pi\in\Pi_{mn}} Y_{\mathbf{t},\pi}\right)^2\right] \\
&= ((mn)!)^2 \cdot \mathop{\mathbb{E}}_{\pi,\pi'\sim\Pi_{mn},\mathbf{t}\sim\mathcal{S}_a}[Y_{\mathbf{t},\pi} \cdot Y_{\mathbf{t},\pi'}] \\
&\overset{(4)}{\leq} ((mn)!)^2 \cdot \left(\sum_{k\geq 1} \frac{q^k}{q^{2n-1}} \cdot \left(\frac{n^2}{(n/2)^{m-2}}\right)^{\frac{k-1}{2}}\right) \\
&= ((mn)!)^2 \cdot \frac{1}{q^{2(n-1)}} \cdot \left(1 + \sum_{k\geq 2} q^{k-1} \cdot \left(\frac{n^2}{(n/2)^{m-2}}\right)^{\frac{k-1}{2}}\right)
\end{aligned}$$

$$= \mu^2 \cdot \left(1 + \sum_{k \geq 2} \left(\frac{(qn)^2}{(n/2)^{m-2}}\right)^{\frac{k-1}{2}}\right).$$

Now, let $p = \left(\frac{(qn)^2}{(n/2)^{m-2}}\right)^{\frac{1}{2}}$. If $m \geq 4 + 100 \log_{n/2}(q/\gamma))$, then we have $p \leq 0.01\gamma^4$. Plugging this back in the above inequality gives

$$\mathop{\mathbb{E}}_{\mathbf{t} \sim \mathcal{S}_a}[Z_{\mathbf{t}}^2] \leq \mu^2 \left(\frac{1}{1-p}\right) \leq \mu^2 \left(\frac{1}{1 - 0.01\gamma^4}\right) \leq \mu^2(1 + 0.02\gamma^4).$$

In other words, we have $\mathrm{Var}_{\mathbf{t} \sim \mathcal{S}_a}(Z_{\mathbf{t}}) \leq (0.2\gamma^2 \cdot \mu)^2$. Hence, by Chebyshev's inequality, we have

$$\Pr_{\mathbf{t} \sim \mathcal{S}_a}[Z_{\mathbf{t}} \leq (1 - 0.5\gamma)\mu] \leq 0.5\gamma. \tag{6}$$

Finally, notice that the statistical distance between $\mathcal{S}_{\mathbf{x}}$ and $\mathcal{S}_a$ is

$$\begin{aligned}
\sum_{\mathbf{t} \in \mathbb{F}_q^{mn}} \max\{f_{\mathcal{S}_a}(\mathbf{t}) - f_{\mathcal{S}_{\mathbf{x}}}(\mathbf{t}), 0\} &= \sum_{\substack{\mathbf{t} \in \mathbb{F}_q^{mn} \\ t_1 + \cdots + t_{mn} = a}} \max\left\{\frac{1}{q^{mn-1}} - \frac{Z_{\mathbf{t}}}{(mn)! \cdot q^{(m-1)n}}, 0\right\} \\
&= \sum_{\substack{\mathbf{t} \in \mathbb{F}_q^{mn} \\ t_1 + \cdots + t_{mn} = a}} f_{\mathcal{S}_a}(\mathbf{t}) \cdot \max\{1 - Z_{\mathbf{t}}/\mu, 0\} \\
&= \mathop{\mathbb{E}}_{\mathbf{t} \sim \mathcal{S}_a}[\max\{1 - Z_{\mathbf{t}}/\mu, 0\}] \\
&\leq \Pr_{\mathbf{t} \sim \mathcal{S}_a}[Z_{\mathbf{t}} \leq (1 - 0.5\gamma)\mu] \cdot 1 + \Pr_{\mathbf{t} \sim \mathcal{S}_a}[Z_{\mathbf{t}} > (1 - 0.5\gamma)\mu] \cdot (0.5\gamma) \\
&\overset{(6)}{\leq} (0.5\gamma) \cdot 1 + 1 \cdot (0.5\gamma) \\
&= \gamma. \qquad \square
\end{aligned}$$

3.2 Relating Moments to Rank of Random Matrices

Having shown how Lemma 1 implies our main theorem (Theorem 3), we now move on to prove Lemma 1 itself. In this subsection, we deal with the first half of the proof by relating the quantity on the left-hand side of (4) to a quantity involving the rank of a certain random matrix.

Warm-Up: (Re-)Computing the First Moment As a first step, let us define below a class of matrices that will be used throughout.

Definition 7. *For every permutation $\pi : [mn] \to [mn]$, let us denote by $\mathbf{A}_\pi \in \mathbb{F}_q^{n \times mn}$ the matrix whose i-th row is the indicator vector for $\pi(\{m(i-1)+1, \ldots, mi\})$. More formally,*

$$(\mathbf{A}_\pi)_{i,j} = \begin{cases} 1 & \text{if } j \in \pi(\{m(i-1)+1, \ldots, mi\}), \\ 0 & \text{otherwise.} \end{cases}$$

Before we describe how these matrices relate to the second moment, let us illustrate their relation to the first moment, by sketching an alternative way to prove Observation 1. To do so, let us rearrange the left-hand side of (3) as $\mathbb{E}_{\mathbf{t}\sim\mathcal{S}_a}[Y_{\mathbf{t},\pi}] = \frac{1}{q^{mn-1}}\sum_{\mathbf{t}\in\mathbb{F}_q^{mn}} Y_{\mathbf{t},\pi}$. Now, observe that $Y_{\mathbf{t},\pi} = 1$ iff $\mathbf{A}_\pi\mathbf{t} = \mathbf{x}$. Since the rows of the matrix $\mathbf{A}_\pi$ have pairwise-disjoint supports, the matrix is always full rank (over $\mathbb{F}_q$), i.e., $\mathrm{rank}(\mathbf{A}_\pi) = n$. This means that the number of values of $\mathbf{t}$ satisfying the aforementioned equation is q^{mn-n}. Plugging this into the above expansion gives $\mathbb{E}_{\mathbf{t}\sim\mathcal{S}_a}[Y_{\mathbf{t},\pi}] = \frac{q^{mn-n}}{q^{mn-1}} = \frac{1}{q^{n-1}}$. Hence, we have rederived (3).

Relating Second Moment to Rank In the previous subsection, we have seen the relation of matrix $\mathbf{A}_\pi$ to the first moment. We will now state such a relation for the second moment. Specifically, we will rephrase the left-hand side of (4) as a quantity involving matrices $\mathbf{A}_\pi$ and $\mathbf{A}_{\pi'}$. To do so, we will need the following additional notations:

Definition 8. *For a pair of permutations $\pi, \pi' : [mn] \to [mn]$, we let $\mathbf{A}_{\pi,\pi'} \in \mathbb{F}_q^{2n\times mn}$ denote the (column-wise) concatenation of $\mathbf{A}_\pi$ and $\mathbf{A}_{\pi'}$, i.e.,*

$$\mathbf{A}_{\pi,\pi'} = \begin{bmatrix}\mathbf{A}_\pi \\ \mathbf{A}_{\pi'}\end{bmatrix}.$$

Furthermore, let[5] *the* rank deficit *of $\mathbf{A}_{\pi,\pi'}$ be* $\mathrm{defc}(\mathbf{A}_{\pi,\pi'}) := 2n - \mathrm{rank}(\mathbf{A}_{\pi,\pi'})$.

Analogous to the relationship between the first moment and $\mathbf{A}_\pi$ seen in the previous subsection, the quantity $\mathbb{E}_{\mathbf{t}\sim\mathcal{S}_a}[Y_{\mathbf{t},\pi}\cdot Y_{\mathbf{t},\pi'}]$ is in fact proportional to the number of solutions to certain linear equations, which is represented by $\mathbf{A}_{\pi,\pi'}$. This allows us to give the bound to the former, as formalized below.

Proposition 1. *For every pair of permutations $\pi, \pi' : [mn] \to [mn]$, we have*

$$\mathop{\mathbb{E}}_{\mathbf{t}\sim\mathcal{S}_a}[Y_{\mathbf{t},\pi}\cdot Y_{\mathbf{t},\pi'}] \le \frac{q^{\mathrm{defc}(\mathbf{A}_{\pi,\pi'})}}{q^{2n-1}}.$$

Proof. First, let us rearrange the left-hand side term as

$$\mathop{\mathbb{E}}_{\mathbf{t}\sim\mathcal{S}_a}[Y_{\mathbf{t},\pi}\cdot Y_{\mathbf{t},\pi'}] = \frac{1}{q^{mn-1}}\sum_{\mathbf{t}\in\mathbb{F}_q^{mn}} Y_{\mathbf{t},\pi}\cdot Y_{\mathbf{t}\cdot\pi'}. \tag{7}$$

Now, notice that $Y_{\mathbf{t},\pi} = 1$ iff $\mathbf{A}_\pi\mathbf{t} = \mathbf{x}$. Similarly, $Y_{\mathbf{t},\pi'} = 1$ iff $\mathbf{A}_{\pi'}\mathbf{t} = \mathbf{x}$. In other words, $Y_{\mathbf{t},\pi}\cdot Y_{\mathbf{t}\cdot\pi'} = 1$ iff

$$\mathbf{A}_{\pi,\pi'}\mathbf{t} = \begin{bmatrix}\mathbf{x}\\ \mathbf{x}\end{bmatrix}.$$

[5] Note that $\mathrm{defc}(\mathbf{A}_{\pi,\pi'})$ is equal to the *corank* of $\mathbf{A}^T_{\pi,\pi'}$.

The number of solutions $\mathbf{t} \in \mathbb{F}_q^{mn}$ to the above equation is at most $q^{mn-\mathrm{rank}(\mathbf{A}_{\pi,\pi'})} = q^{(m-2)n+\mathrm{defc}(\mathbf{A}^T_{\pi,\pi'})}$. Plugging this back into (7), we get

$$\mathop{\mathbb{E}}_{\mathbf{t}\sim\mathcal{S}_a}[Y_{\mathbf{t},\pi} \cdot Y_{\mathbf{t},\pi'}] \leq \frac{1}{q^{mn-1}} \cdot q^{(m-2)n+\mathrm{defc}(\mathbf{A}_{\pi,\pi'})} = \frac{q^{\mathrm{defc}(\mathbf{A}_{\pi,\pi'})}}{q^{2n-1}},$$

as desired. □

3.3 Probabilistic Bound on Rank Deficit of Random Matrices

The final step of our proof is to bound the probability that the rank deficit of $\mathbf{A}_{\pi,\pi'}$ is large. Such a bound is encapsulated in Theorem 4 below. Notice that Proposition 1 and Theorem 4 immediately yield Lemma 1.

Theorem 4. *For all $m \geq 3$ and $k \in \mathbb{N}$, we have*

$$\Pr_{\pi,\pi'\sim\Pi_{mn}}[\mathrm{defc}(\mathbf{A}_{\pi,\pi'}) \geq k] \leq \left(\frac{n^2}{(n/2)^{m-2}}\right)^{\frac{k-1}{2}}.$$

Characterization of Rank Deficit via Matching Partitions. To prove Theorem 4, we first give a "compact" and convenient characterization of the rank deficit of $\mathbf{A}_{\pi,\pi'}$. In order to do this, we need several additional notations: we say that a partition $S_1 \sqcup \cdots \sqcup S_k = U$ of a universe U is *non-empty* if $S_1, \ldots, S_k \neq \emptyset$. Moreover, for a set $S \subseteq [n]$, we use $S^{\to m} \subseteq [mn]$ to denote the set $\cup_{i\in S}\{m(i-1)+1, \ldots, mi\}$. Finally, we need the following definition of *matching partitions*.

Definition 9. *Let π, π' be any pair of permutations of $[mn]$. A pair of non-empty partitions $S_1 \sqcup \cdots \sqcup S_k = [n]$ and $S'_1 \sqcup \cdots \sqcup S'_k = [n]$ is said to* match with respect to π, π' *iff*

$$\pi\left(S_j^{\to m}\right) = \pi'\left((S'_j)^{\to m}\right) \tag{8}$$

for all $j \in [k]$. When π, π' are clear from the context, we may omit "with respect to π, π'" from the terminology.

Condition (8) might look a bit mysterious at first glance. However, there is a very simple equivalent condition in terms of the matrices $\mathbf{A}_\pi, \mathbf{A}_{\pi'}$: $S_1 \sqcup \cdots \sqcup S_k = [n]$ and $S'_1 \sqcup \cdots \sqcup S'_k = [n]$ match iff the sum of rows $i \in S_j$ of $\mathbf{A}_\pi$ coincides with the sum of rows $i' \in S'_j$ of $\mathbf{A}_{\pi'}$, i.e., $\sum_{i\in S_j}(\mathbf{A}_\pi)_i = \sum_{i'\in S'_j}(\mathbf{A}_{\pi'})_{i'}$.

An easy-to-use equivalence of $\mathrm{defc}(\mathbf{A}_{\pi,\pi'}) = k$ is that a pair of matching partitions $S_1 \sqcup \cdots \sqcup S_k = [n]$ and $S'_1 \sqcup \cdots \sqcup S'_k = [n]$ exists. We only use one direction of this relation, which we prove below.

Lemma 2. *For any permutations $\pi, \pi' : [mn] \to [mn]$, if $\mathrm{defc}(\mathbf{A}_{\pi,\pi'}) \geq k$, then there exists a pair of matching partitions $S_1 \sqcup \cdots \sqcup S_k = [n]$ and $S'_1 \sqcup \cdots \sqcup S'_k = [n]$.*

Proof. We will prove the contrapositive. Let $\pi, \pi' : [mn] \to [mn]$ be any permutations, and suppose that there is no pair of matching partitions $S_1 \sqcup \cdots \sqcup S_k = [n]$ and $S'_1 \sqcup \cdots \sqcup S'_k = [n]$. We will show that $\mathrm{defc}(\mathbf{A}_{\pi,\pi'}) < k$, or equivalently $\mathrm{rank}(\mathbf{A}_{\pi,\pi'}) > 2n - k$.

Consider any pair of matching partitions[6] $S_1 \sqcup \cdots \sqcup S_t = [n]$ and $S'_1 \sqcup \cdots \sqcup S'_t = [n]$ that maximizes the number of parts t. From our assumption, we must have $t < k$.

For every part $j \in [t]$, let us pick an arbitrary element $i_j \in S_j$. Consider all rows of $\mathbf{A}_{\pi,\pi'}$, except the i_j-th rows for all $j \in [t]$ (i.e. $\{(\mathbf{A}_{\pi,\pi'})_i\}_{i \notin \{i_1,\ldots,i_t\}}$). We claim that these rows are linearly independent. Before we prove this, note that this imply that the rank of $\mathbf{A}_{\pi,\pi'}$ is at least $2n - t > 2n - k$, which would complete our proof.

We now move on to prove the linear independence of $\{(\mathbf{A}_{\pi,\pi'})_i\}_{i \notin \{i_1,\ldots,i_t\}}$. Suppose for the sake of contradiction that these rows are not linearly independent. Since the matrix $\mathbf{A}_{\pi,\pi'}$ is simply a concatenation of $\mathbf{A}_\pi$ and $\mathbf{A}_{\pi'}$, we have that $\{(\mathbf{A}_{\pi,\pi'})_i\}_{i \notin \{i_1,\ldots,i_t\}} = \{(\mathbf{A}_\pi)_i\}_{i \in [n] \setminus \{i_1,\ldots,i_t\}} \cup \{(\mathbf{A}_{\pi'})_{i'}\}_{i' \in [n]}$. The linear dependency of these rows mean that there exists a non-zero vector of coefficients $(c_1, \ldots, c_n, c'_1, \ldots, c'_n) \in \mathbb{F}_q^{2n}$ with $c_{i_1} = \cdots = c_{i_t} = 0$ such that

$$\mathbf{0} = \sum_{i \in [n]} c_i \cdot (\mathbf{A}_\pi)_i + \sum_{i' \in [n]} c'_{i'} \cdot (\mathbf{A}_{\pi'})_{i'}. \tag{9}$$

Since the rows of $\mathbf{A}_{\pi'}$ are linearly independent, there must exist $i^* \in [n]$ such that $c_{i^*} \neq 0$. Let $j \in [t]$ denote the index of the partition to which i^* belongs, i.e., $i^* \in S_j$. For notational convenience, we will assume, without loss of generality, that $j = t$.

Let $P_t : \mathbb{F}_q^{mn} \to \mathbb{F}_q^{(S_t^{\to m})}$ denote the projection operator that sends a vector $(v_\ell)_{\ell \in [mn]}$ to its restriction on coordinates in $S_t^{\to m}$, i.e., $(v_\ell)_{\ell \in S_t^{\to m}}$. Observe that $P_t((\mathbf{A}_\pi)_i)$ is non-zero iff $i \in S_t$ and $P_t((\mathbf{A}_{\pi'})_{i'})$ is non-zero iff $i' \in S'_t$. Thus, by taking P_t on both sides of (9), we have

$$\mathbf{0} = \sum_{i \in S_t} c_i \cdot P_t((\mathbf{A}_\pi)_i) + \sum_{i' \in S'_t} c_{i'} \cdot P_t((\mathbf{A}_\pi)_{i'}) \tag{10}$$

Now, let $T = \{i \in S_t \mid c_i \neq 0\}$ and $T' = \{i' \in S'_t \mid c_{i'} \neq 0\}$. Notice that $\mathrm{supp}\left(\sum_{i \in S_t} c_i \cdot P_t((\mathbf{A}_\pi)_i)\right) = \pi(T^{\to m})$ and $\mathrm{supp}\left(\sum_{i' \in S'_t} c_{i'} \cdot P_t((\mathbf{A}_\pi)_{i'})\right) = \pi'((T')^{\to m})$. Hence, from (10), we have

$$\pi(T^{\to m}) = \pi'((T')^{\to m}). \tag{11}$$

Consider the pair of partitions $S_1 \sqcup \cdots S_{t-1} \sqcup T \sqcup (S_t \setminus T) = [n]$ and $S'_1 \sqcup \cdots S'_{t-1} \sqcup T' \sqcup (S'_t \setminus T') = [n]$. From the definition of T, we must have $T \neq \emptyset$ because i^* belongs to T, and $(S_t \setminus T) \neq \emptyset$ becase i_t does not belong to T. From this and (11), these partitions are non-empty and they match. However, these matching partitions have $t + 1$ parts, which contradicts the maximality of the number of parts of $S_1 \sqcup \cdots \sqcup S_t$ and $S'_1 \sqcup \cdots \sqcup S'_t$. This concludes our proof. □

[6] Note that at least one matching partition always exists: $S_1 = [n] = S'_1$.

Proof of Theorem 4 With the characterization from the previous subsection ready, we can now easily prove our main theorem of this section (Theorem 4). We will also use two simple inequalities regarding the multinomial coefficients stated below. For completeness, we provide their proofs in the appendix.

Fact 1. *For every $a_1, \dots, a_k, a'_1, \dots, a'_k \in \mathbb{N}$, we have*

$$\binom{a_1 + \cdots + a_k + a'_1 + \cdots + a'_k}{a_1 + a'_1, \dots, a_k + a'_k} \geq \binom{a_1 + \cdots + a_k}{a_1, \dots, a_k} \cdot \binom{a'_1 + \cdots + a'_k}{a'_1, \dots, a'_k}$$

Fact 2. *For every $k \in \mathbb{N}$ and $a_1, \dots, a_k \in \mathbb{N}$, we have*

$$\binom{a_1 + \cdots + a_k}{a_1, \dots, a_k} \geq \left(\frac{a_1 + \cdots + a_k}{2}\right)^{\lfloor k/2 \rfloor}$$

Proof of Theorem 4. Let us fix a pair of non-empty partitions $S_1 \sqcup \cdots \sqcup S_k = [n]$ and $S'_1 \sqcup \cdots \sqcup S'_k = [n]$ such that[7] $|S_i| = |S'_i|$ for all $i \in [k]$. Notice that, when we pick $\pi : [mn] \to [mn]$ uniformly at random, $(\pi(S_1^{\to m}), \cdots, \pi(S_k^{\to m}))$ is simply a random partition of $[mn]$ into subsets of size $m|S_1|, \dots, m|S_k|$. Hence, the probability that these partitions match is equal to

$$\frac{1}{\binom{mn}{m|S_1|, \dots, m|S_k|}}.$$

Hence, by evoking Lemma 2 and taking union bound over all pairs of partitions $S_1 \sqcup \cdots \sqcup S_k = [n]$ and $S'_1 \sqcup \cdots \sqcup S'_k = [n]$, we have

$$\begin{aligned}
\Pr_{\pi, \pi' \sim \Pi_{mn}}[\mathrm{defc}(\mathbf{A}^T_{\pi,\pi'}) \geq k] &\leq \sum_{\substack{S_1 \sqcup \cdots \sqcup S_k = [n], S'_1 \sqcup \cdots \sqcup S'_k = [n] \\ |S_1| = |S'_1| > 0, \dots, |S_k| = |S'_k| > 0}} \frac{1}{\binom{mn}{m|S_1|, \dots, m|S_k|}} \\
&= \sum_{\substack{a_1, \dots, a_k \in \mathbb{N} \\ a_1 + \cdots + a_k = n}} \sum_{\substack{S_1 \sqcup \cdots \sqcup S_k = [n], S'_1 \sqcup \cdots \sqcup S'_k = [n] \\ |S_1| = |S'_1| = a_1, \dots, |S_k| = |S'_k| = a_k}} \frac{1}{\binom{mn}{ma_1, \dots, ma_k}} \\
&= \sum_{\substack{a_1, \dots, a_k \in \mathbb{N} \\ a_1 + \cdots + a_k = n}} \frac{\binom{n}{a_1, \dots, a_k}^2}{\binom{mn}{ma_1, \dots, ma_k}} \\
(\text{Fact 1}) &\leq \sum_{\substack{a_1, \dots, a_k \in \mathbb{N} \\ a_1 + \cdots + a_k = n}} \frac{1}{\binom{n}{a_1, \dots, a_k}^{(m-2)}} \\
(\text{Fact 2}) &\leq \sum_{\substack{a_1, \dots, a_k \in \mathbb{N} \\ a_1 + \cdots + a_k = n}} \frac{1}{(n/2)^{(m-2) \cdot \lfloor k/2 \rfloor}} \\
&\leq \frac{n^{k-1}}{(n/2)^{(m-2) \cdot \lfloor k/2 \rfloor}}
\end{aligned}$$

[7] We may assume that $|S_i| = |S'_i|$; otherwise, $\pi(S_i^{\to m})$ and $\pi'((S'_i)^{\to m})$ are obviously not equal and hence $S_1 \sqcup \cdots \sqcup S_k = [n]$ and $S'_1 \sqcup \cdots \sqcup S'_k = [n]$ do not match.

$$\leq \left(\frac{n^2}{(n/2)^{m-2}}\right)^{\frac{k-1}{2}} \qquad \square$$

4 Lower Bound Proofs

In this section, we prove our lower bound on the number of messages (Theorem 2), which is a direct consequence of the following two theorems:

Theorem 5. *Suppose $\sigma \geq 1$. Then, for any σ-secure n-party aggregation protocol over $\mathbb{F}_q$ in which each party sends m messages, we have $m = \Omega(\log_n q)$.*

Theorem 6. *For any σ-secure n-party aggregation protocol over $\mathbb{F}_q$ in which each party sends m messages, we have $m = \Omega\left(\frac{\sigma}{\log(\sigma n)}\right)$.*

We prove Theorem 5 in Section 4.1, while we prove Theorem 6 in Section 4.2. Before we proceed to the proofs, let us start by proving the following fact that will be used in both proofs: the output of the encoder on a party's input must uniquely determine the input held by the party.

Lemma 3. *For any n-party aggregation protocol $\mathcal{P}$ with encoder $\mathsf{Enc} : \mathbb{F}_q \to [\ell]^m$, we have that for any $x, x' \in \mathbb{F}_q$ with $x \neq x'$, the distributions $\mathcal{E}_x^{\mathsf{Enc}}$ and $\mathcal{E}_{x'}^{\mathsf{Enc}}$ have disjoint supports.*

As a consequence, for any output vector $\mathbf{y} \in [\ell]^{nm}$, there exists at most one $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{F}_q^n$ such that $\mathbf{y}$ is a possible output $(\mathsf{Enc}_{x_1}, \mathsf{Enc}_{x_2}, \ldots, \mathsf{Enc}_{x_n})$.

Proof. For the sake of contradiction, suppose there exist $x, x' \in \mathbb{F}_q$ with $x \neq x'$ such that $\mathcal{E}_x^{\mathsf{Enc}}$ and $\mathcal{E}_{x'}^{\mathsf{Enc}}$ have a common element in the support, say $\mathbf{z}$. Then, let $\mathbf{z}' \in [\ell]^m$ be an element in the support of $\mathcal{E}_0^{\mathsf{Enc}}$. Note that it follows that $(\mathbf{z}, \underbrace{\mathbf{z}', \mathbf{z}', \ldots, \mathbf{z}'}_{n-1})$ is a possible output of inputs $(x, \mathbf{0}^{n-1})$ and $(x', \mathbf{0}^{n-1})$, which means that the analyzer cannot uniquely determine the parties' inputs from the output, thereby contradicting the correctness of the protocol. This completes the proof. $\square$

4.1 Field-Dependent Bound

We now present the proof of Theorem 5. Recall from Section 1.4 that $\mathcal{B}_s$ is defined as $\{\mathbf{x} \in \mathbb{F}_q^n \mid \sum_i x_i = s\}$. The key technical lemma is the following.

Lemma 4. *For each $s \in \mathbb{F}_q$ and every n-user one-round aggregation protocol $\mathcal{P}$ in the anonymized model with encoder $\mathsf{Enc} : \mathbb{F}_q \to [\ell]^m$, there exists a pair of inputs $\mathbf{x}, \mathbf{x}' \in \mathcal{B}_s$ such that $\mathrm{SD}\left(\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}, \mathcal{S}_{\mathbf{x}'}^{\mathsf{Enc}}\right) \geq 1 - n^{nm}/q^{n-1}$.*

Throughout this subsection, let us fix $s \in \mathbb{F}_q$. Before proving Lemma 4, we first define some notation. For every possible shuffler output vector $\mathbf{y}$ and input $\mathbf{x} \in \mathcal{B}_s$, let $p_{\mathbf{x},\mathbf{y}}$ denote the probability that on input $\mathbf{x}$ the encoder outputs $\mathbf{y}$, i.e., $\Pr_{Y \sim \mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}}[Y = \mathbf{y}]$. Moreover, let $\mathrm{Inv}_{\mathbf{y}} = \{\mathbf{x} \in \mathcal{B}_s \mid p_{\mathbf{x},\mathbf{y}} > 0\}$ denote the set of sum-s inputs that are possible given that the output is $\mathbf{y}$.

Lemma 5. $|\mathrm{Inv}_{\mathbf{y}}| \leq n^{nm}$.

Proof. Suppose $\mathbf{y}$ is an output vector consisting of nm messages with $|\mathrm{Inv}_{\mathbf{y}}| > 0$. Consider a function $g : [nm] \to [n]$ that associates each of the mn messages to a single party. Note that $\mathbf{y}$ and g uniquely identify the set of messages Y_i sent by each party i. In turn, Y_i must correspond to a unique input x_i to party i by Lemma 3. Then, it follows that $\mathbf{y}$ and g can determine at most one input $\mathbf{x} \in \mathrm{Inv}_{\mathbf{y}}$. Since there are at most n^{nm} valid functions g, the desired bound on $|\mathrm{Inv}_{\mathbf{y}}|$ follows. □

Let $p_{\mathbf{y}} = \sum_{\mathbf{x} \in \mathcal{B}_s} p_{\mathbf{x},\mathbf{y}}$, and define $d_{\mathbf{y}} = \frac{1}{q^{2n-2}} \sum_{\mathbf{x} \in \mathcal{B}_s} \sum_{\mathbf{x}' \in \mathcal{B}_s} |p_{\mathbf{x},\mathbf{y}} - p_{\mathbf{x}',\mathbf{y}}|$ as the average difference between probabilities $p_{\mathbf{x},\mathbf{y}}$ and $p_{\mathbf{x}',\mathbf{y}}$ over all pairs of inputs $\mathbf{x}, \mathbf{x}'$ with sum s. Then, we have the following lemma.

Lemma 6. $d_{\mathbf{y}} \geq 2\left(1 - \frac{n^{nm}}{q^{n-1}}\right) p_{\mathbf{y}} / q^{n-1}$.

Proof. We have

$$\begin{aligned} q^{2n-2} d_{\mathbf{y}} &\geq 2 \sum_{\mathbf{x} \in \mathrm{Inv}_{\mathbf{y}}} \sum_{\mathbf{x}' \in \mathcal{B}_s \setminus \mathrm{Inv}_{\mathbf{y}}} |p_{\mathbf{x},\mathbf{y}} - \mathbf{0}| \\ &= 2\,|\mathcal{B}_s \setminus \mathrm{Inv}_{\mathbf{y}}| \sum_{\mathbf{x} \in \mathrm{Inv}_{\mathbf{y}}} p_{\mathbf{x},\mathbf{y}} \\ &= 2\left(q^{n-1} - |\mathrm{Inv}_{\mathbf{y}}|\right) p_{\mathbf{y}} \\ (\text{Lemma 5}) &\geq 2\left(q^{n-1} - n^{nm}\right) p_{\mathbf{y}}. \end{aligned}$$

□

We now prove Lemma 4.

Proof of Lemma 4. We will in fact show the stronger statement that the (scaled) *average* statistical distance for pairs of inputs in $\mathcal{B}_s$ is lower bounded by $1 - n^{nm}/q^{n-1}$, i.e.,

$$d_{\mathrm{avg}} \geq 1 - \frac{n^{nm}}{q^{n-1}},$$

where

$$d_{\mathrm{avg}} = \frac{1}{q^{2n-2}} \sum_{\mathbf{x} \in \mathcal{B}_s} \sum_{\mathbf{x}' \in \mathcal{B}_s} \mathrm{SD}\left(\mathcal{S}^{\mathsf{Enc}}_{\mathbf{x}}, \mathcal{S}^{\mathsf{Enc}}_{\mathbf{x}'}\right). \tag{12}$$

Note that by Lemma 6, we have

$$\begin{aligned} d_{\mathrm{avg}} &= \sum_{\mathbf{y}} \frac{d_{\mathbf{y}}}{2} \\ &\geq \frac{1}{q^{n-1}} \left(1 - \frac{n^{nm}}{q^{n-1}}\right) \sum_{\mathbf{y}} p_{\mathbf{y}} \\ &\geq \frac{1}{q^{n-1}} \left(1 - \frac{n^{nm}}{q^{n-1}}\right) \sum_{\mathbf{y}} \sum_{\mathbf{x} \in \mathcal{B}_s} p_{\mathbf{x},\mathbf{y}} \end{aligned}$$

$$= 1 - \frac{n^{nm}}{q^{n-1}},$$

where the last line follows from the fact that $|\mathcal{B}_s| = q^{n-1}$. To conclude, note that it follows that at least one of the summands in (12) must be at least $1 - \frac{n^{nm}}{q^{n-1}}$, as desired. □

Theorem 5 now follows easily from Lemma 4.

Proof of Theorem 5. Suppose $\mathcal{P}$ is such a σ-secure n-party aggregation protocol with encoder $\mathsf{Enc} : \mathbb{F}_q \to [\ell]^m$. Then, choose an arbitrary $s \in \mathbb{F}_q$. Note that by Lemma 4, there exist $\mathbf{x}, \mathbf{x}' \in \mathcal{B}_s$ such that $2^{-\sigma} \geq \mathrm{SD}\left(\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}, \mathcal{S}_{\mathbf{x}'}^{\mathsf{Enc}}\right) \geq 1 - \frac{n^{nm}}{q^{n-1}}$. Thus, if $\sigma \geq 1$, it follows that $m = \Omega(\log_n q)$, as desired. □

4.2 Security-Dependent Bound

We now turn to the proof of Theorem 6, which follows from the next theorem.

Theorem 7. *Let* Enc *be the encoder of any summation protocol for* $n > 2$ *parties with* m *messages sent per party. Then, there exists a vector* $\mathbf{x} \in \mathcal{B}_0$ *such that the statistical distance between* $\mathcal{S}_{\mathbf{0}}^{\mathsf{Enc}}$ *and* $\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}$ *is at least* $\frac{1}{(10nm)^{5m}}$.

It is not hard to see that Theorem 6 follows from Theorem 7:

Proof of Theorem 6. Simply note that by Theorem 7 and the definition of σ-security, we can find $\mathbf{x} \in \mathcal{B}_0$ such that $2^{-\sigma} \geq \mathrm{SD}\left(\mathcal{S}_{\mathbf{0}}^{\mathsf{Enc}}, \mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}\right) \geq \frac{1}{(10nm)^{5m}}$, which immediately implies that $m = \Omega\left(\frac{\sigma}{\log(\sigma n)}\right)$, as desired. □

Henceforth, we focus on proving Theorem 7.

Warm-up: Proof of Theorem 7 for Ishai et al.'s protocol. Before we prove Theorem 7 for the general case, let us sketch a proof specific to Ishai et al.'s protocol. The input vector $\mathbf{x}$ we will use is simply $\mathbf{x} = (1, \cdots, 1, -(n-1))$.

To lower bound $\mathrm{SD}(\mathcal{S}_{\mathbf{0}}, \mathcal{S}_{\mathbf{x}})$, we give a "distinguisher" $\mathcal{A}$ that takes in the output $(y_1, \ldots, y_{\pi(mn)})$ of the shuffler and outputs either 1 (i.e. "accept") or 0 (i.e. "reject"). Its key property will be that the probability that $\mathcal{A}$ accepts when $(y_{\pi(1)}, \ldots, y_{\pi(mn)}) \sim \mathcal{S}_{\mathbf{0}}$ is more than that of when $(y_{\pi(1)}, \ldots, y_{\pi(mn)}) \sim \mathcal{S}_{\mathbf{x}}$ by an additive factor of $\frac{1}{(en)^m}$. This immediately implies that the distributions $\mathcal{S}_{\mathbf{0}}$ and $\mathcal{S}_{\mathbf{x}}$ are at a statistical distance of at least $\frac{1}{(en)^m}$ as well. (Note that this bound is slightly better than the one in Theorem 7.)

The distinguisher $\mathcal{A}$ is incredibly simple here: $\mathcal{A}$ accepts iff $y_{\pi(1)} + \cdots + y_{\pi(m)} = 0$. To see that it satisfies the claim property, observe that, when $\pi(1), \ldots, \pi(m)$ not all come from the same party, $y_{\pi(1)} + \cdots + y_{\pi(m)}$ is simply a random number in $\mathbb{F}_q$, meaning that $\mathcal{A}$ accepts with probability $1/q$ (in both distributions). On the other hand, when $\pi(1), \ldots, \pi(m)$ come from the same party, $y_{\pi(1)} + \cdots + y_{\pi(m)}$ is always zero in the distribution $\mathcal{S}_{\mathbf{0}}$ and hence $\mathcal{A}$

always accept. For the distribution $\mathcal{S}_{\mathbf{x}}$, if $\pi(1), \ldots, \pi(m)$ comes from the same party $i \neq n$, then the sum $y_{\pi(1)} + \cdots + y_{\pi(m)}$ is always one and hence $\mathcal{A}$ rejects. Thus, the probability that $\mathcal{A}$ accepts in the former distribution is more than that of the latter by an additive factor of $\frac{n-1}{\binom{nm}{m}} \geq \frac{1}{(en)^m}$. (The -1 factor corresponds to the case where $p(1), \cdots, p(m)$ comes from party $i = n$; here $\mathcal{A}$ might accept if $-(n-1) = 0$ in $\mathbb{F}_q$.) This concludes the proof sketch.

From Ishai et al.'s protocol to general protocols. Having sketched the argument for Ishai et al.'s protocol, one might wonder whether the same approach would work for general protocols. In particular, here instead of checking if $y_{\pi(1)} + \cdots + y_{\pi(m)} = 0$, we would check whether $y_{\pi(1)}, \ldots, y_{\pi(m)}$ is a valid output of the encoder when the input is zero. Now, the statement for when $\pi(1), \ldots, \pi(m)$ comes from the same party remains true. However, the issue is that, when $\pi(1), \ldots, \pi(m)$ do not all come from the same party, it is not necessarily true that the acceptance probability of $\mathcal{A}$ would be the same for both distributions.

To avoid having these "cross terms" affect the probability of acceptance of $\mathcal{A}$ too much, we pick the smallest integer t such that the "t-message marginals" (defined formally below) of $\mathcal{E}_0^{\mathsf{Enc}}$ and $\mathcal{E}_1^{\mathsf{Enc}}$ differ "substantially". Then, we modify $\mathcal{A}$ so that it performs an analogous check on $y_{\pi(1)}, \ldots, y_{\pi(t)}$ (instead of $y_{\pi(1)}, \ldots, y_{\pi(m)}$ as before). Once again, we will have that, if $\pi(1), \ldots, \pi(t)$ corresponds to the same party, then the probability that $\mathcal{A}$ accepts differs significantly between the two cases. On the other hand, due to the minimality of t, we can also argue that, when $\pi(1), \ldots, \pi(t)$ are not all from the same parties (i.e. "cross terms"), the difference is small. Hence, the former case would dominate and we can get a lower bound on the difference as desired. This is roughly the approach we take in the proof of Theorem 7 below. There are subtle points we have to change in the actual proof below. For instance, we cannot simply use the input $(1, \cdots, 1, -(n-1))$ as in the case of Ishai et al. protocol because, if the t-marginal of $\mathcal{E}_{-(n-1)}^{\mathsf{Enc}}$ deviates from $\mathcal{E}_0^{\mathsf{Enc}}$ more substantially than that of $\mathcal{E}_1^{\mathsf{Enc}}$, then this could affect the acceptance probability by a lot. Hence, in the actual proof, we instead pick x^* that minimizes the value of such t among all numbers in $\mathbb{F}_q$, and use the input vector $\mathbf{x} = (x^*, \ldots, x^*, -(n-1)x^*)$.

Additional Notation and Observation. To formally prove Theorem 7 in the general form, we need to formally define the notion of t-marginal. For a distribution $\mathcal{D}$ supported on $[\ell]^m$ and a positive integer $t \leq m$, its *t-marginal*, denoted by $\mathcal{D}|_t$, supported on $[\ell]^t$ is simply the marginal of $\mathcal{D}$ on the first t-coordinates; more formally, for all $\mathbf{y} \in [\ell]^t$, we have

$$\Pr_{Y \sim \mathcal{D}|_t}[Y = \mathbf{y}] = \sum_{y_{t+1}, \ldots, y_m \in [\ell]} \Pr_{Y \sim \mathcal{D}}[Y = \mathbf{y} \circ (y_{t+1}, \ldots, y_m)].$$

An observation that will simplify our proof is that we may assume w.l.o.g. that the distribution $\mathcal{E}_x^{\mathsf{Enc}}$ for every $x \in \mathbb{F}_q$ is permutation invariant, i.e., that

for any $\pi : [m] \to [m]$ and any $\mathbf{y} \in [\ell]^m$, we have

$$\Pr_{Y \sim \mathcal{E}_x^{\mathsf{Enc}}}[Y = \mathbf{y}] = \Pr_{Y \sim \mathcal{E}_x^{\mathsf{Enc}}}[Y = \pi(\mathbf{y})].$$

This is because we may apply a random permutation to the encoding Enc_x before sending it to the shuffler, which does not change the distribution $\mathcal{S}^x_{\mathsf{Enc}}$. Notice that our observation implies that $\mathcal{E}_x^{\mathsf{Enc}}|_t$ is also permutation invariant.

Proof of Theorem 7. Let $t \leq m$ be the smallest positive integer such that $\max_{x \in \mathbb{F}_q} \mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|_t, \mathcal{E}_x^{\mathsf{Enc}}|_t)$ is at least $\frac{1}{(10nm)^{4(m-t)}}$. Note that such t always exist because the requirement holds for $t = m$, at which $\mathcal{E}_0^{\mathsf{Enc}}|_t = \mathcal{E}_0^{\mathsf{Enc}}$ and $\mathcal{E}_1^{\mathsf{Enc}}|_t = \mathcal{E}_1^{\mathsf{Enc}}$ have statistical distance 1 (as their supports are disjoint due to Lemma 3).

For t as defined above, let $x^* = \mathrm{argmax}_{x \in \mathbb{F}_q} \mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|t, \mathcal{E}_x^{\mathsf{Enc}}|_t)$ and let us defined H as the set of elements of $[\ell]^t$ whose probability under $\mathcal{E}_0^{\mathsf{Enc}}|_t$ is higher than under $\mathcal{E}_{x^*}^{\mathsf{Enc}}|_t$. More formally, $H = \{\mathbf{y} \in [\ell]^t : \mathcal{E}_0^{\mathsf{Enc}}|_t(\mathbf{y}) > \mathcal{E}_{x^*}^{\mathsf{Enc}}|_t(\mathbf{y})\}$. By definition of statistical distance, we have

$$\Pr_{\mathbf{y} \in \mathcal{E}_0^{\mathsf{Enc}}|_t}[\mathbf{y} \in H] - \Pr_{\mathbf{y} \in \mathcal{E}_{x^*}^{\mathsf{Enc}}|_t}[\mathbf{y} \in H] = SD(\mathcal{E}_0^{\mathsf{Enc}}|_t, \mathcal{E}_{x^*}^{\mathsf{Enc}}|_t) \geq \frac{1}{(10nm)^{4(m-t)}}, \quad (13)$$

where the inequality follows from our choice of t.

Let $\mathbf{x} = (x^*, \ldots, x^*, -(n-1)x^*)$; clearly, $\mathbf{x} \in \mathcal{B}_0$ as desired. We next give a distinguisher for the distributions $\mathcal{S}_{\mathbf{0}}^{\mathsf{Enc}}$ and $\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}$. The distinguisher $\mathcal{A}$ takes in the permuted output $(y_{\pi(1)}, \ldots, y_{\pi(nm)})$. It returns one (i.e., "accept") if $(y_{\pi(1)}, \ldots, y_{\pi(t)})$ belongs to H and it returns zero (i.e., "reject") otherwise.

We will show that the probability that $\mathcal{A}$ accepts on $\mathcal{S}_{\mathbf{0}}^{\mathsf{Enc}}$ is more than the probability that it accepts on $\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}$ by at least $\frac{1}{(10nm)^{5m}}$, which implies that the statistical distance between $\mathcal{S}_{\mathbf{0}}^{\mathsf{Enc}}$ and $\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}$ is also at least $\frac{1}{(10nm)^{5m}}$ as desired.

To argue about the acceptance probability of $\mathcal{A}$, it is worth noting that there are two sources of randomness here: the output $\mathbf{y}$ (sampled from $\mathcal{E}_{\mathbf{0}}^{\mathsf{Enc}}$ or $\mathcal{E}_{\mathbf{x}}^{\mathsf{Enc}}$) and the permutation π. More formally, we may write the probability that $\mathcal{A}$ accepts on $\mathcal{S}_{\mathbf{0}}^{\mathsf{Enc}}$ and that on $\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}$ as $\Pr_{\pi \sim \Pi_{mn}, \mathbf{y} \sim \mathcal{E}_{\mathbf{0}}^{\mathsf{Enc}}}[\mathcal{A}(\pi(y)) = 1]$ and $\Pr_{\pi \sim \Pi_{mn}, \mathbf{y} \sim \mathcal{E}_{\mathbf{x}}^{\mathsf{Enc}}}[\mathcal{A}(\pi(y)) = 1]$ respectively. Hence, the difference between the probability that $\mathcal{A}$ accepts on $\mathcal{S}_{\mathbf{0}}^{\mathsf{Enc}}$ and that on $\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}$ is

$$\begin{aligned}
&\Pr_{\pi \sim \Pi_{mn}, \mathbf{y} \sim \mathcal{S}_{\mathbf{0}}^{\mathsf{Enc}}}[\mathcal{A}(\pi(y)) = 1] - \Pr_{\pi \sim \Pi_{mn}, \mathbf{y} \sim \mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}}[\mathcal{A}(\pi(y)) = 1] \\
&= \mathop{\mathbb{E}}_{\pi \sim \Pi_{mn}} \left[\Pr_{\mathbf{y} \sim \mathcal{S}_{\mathbf{0}}^{\mathsf{Enc}}}[\mathcal{A}(\pi(y)) = 1] - \Pr_{\mathbf{y} \sim \mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}}[\mathcal{A}(\pi(y)) = 1] \right].
\end{aligned}$$

For brevity, let us define Δ_π as

$$\Delta_\pi := \Pr_{\mathbf{y} \sim \mathcal{S}_{\mathbf{0}}^{\mathsf{Enc}}}[\mathcal{A}(\pi(y)) = 1] - \Pr_{\mathbf{y} \sim \mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}}[\mathcal{A}(\pi(y)) = 1].$$

Note that the quantity we would like to lower bound is now simply $\mathbb{E}_\pi[\Delta_\pi]$.

For each party $i \in \{1, \ldots, n\}$ and any permutation $\pi : [mn] \to [mn]$, we use U_π^i to denote $\{\pi(1), \ldots, \pi(t)\} \cap \{m(i-1)+1, \ldots, mi\}$. Furthermore, we define *the largest number of messages from a single party* for a permutation π as $C_\pi := \max_{i=1,\ldots,n} |U_\pi^i|$.

In the next part of the proof, we classify π into three categories, as listed below. For each category, we prove either a lower or an upper bound on Δ_π and the probability that a random permutation falls into that category.

I. $C_\pi = t$ and $|U_\pi^n| \neq t$. In other words, all of $\{\pi(1), \ldots, \pi(t)\}$ correspond to a single party and that party is not the last party.
II. $C_\pi = t$ and $|U_\pi^n| = t$. In other words, all of $\{\pi(1), \ldots, \pi(t)\}$ correspond to the last party n.
III. $C_\pi < t$. Not all of $\pi(1), \ldots, \pi(t)$ comes from the same party.

We will show that for category I permutations, Δ_π is large (Lemma 11) and the probability that a random permutation belongs to this category is not too small (Lemma 8). For both categories II and III, we show that $|\Delta_\pi|$ is small (Lemmas 9 and 11) and the probabilities that a random permutation belongs to each of these two categories are not too large (Lemmas 10 and 12).

These quantitative bounds are such that the first category dominates $\mathbb{E}_\pi[\Delta_\pi]$, meaning that we get a lower bound on this expectation as desired; this is done at the very end of the proof.

Category I: $C_\pi = t$ and $|U_\pi^n| \neq t$.

We now consider the first case: when $\{\pi(1), \ldots, \pi(t)\}$ corresponds to a single party $i \neq n$. In this case, Δ_π is exactly equal to the statistical distance between $\mathcal{E}_0^{\mathsf{Enc}}$ and $\mathcal{E}_{x^*}^{\mathsf{Enc}}$ (which we know from (13) to be large):

Lemma 7. *For any π such that $C_\pi = t$ and $|U_\pi^n| \neq t$, we have $\Delta_\pi = \mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|_t, \mathcal{E}_{x^*}^{\mathsf{Enc}}|_t)$.*

Proof. Let $i \in \{1, \ldots, n\}$ be the party such that $|U_\pi^i| = C_\pi = t$. When $\mathbf{y}$ is drawn from $\mathcal{E}_{\mathbf{x}}^{\mathsf{Enc}}$ (respectively $\mathcal{E}_{\mathbf{0}}^{\mathsf{Enc}}$), $\{\pi(1), \cdots, \pi(t)\} \subseteq \{m(i-1)+1, \ldots, mi\}$, it is the case that $(y_{\pi(1)}, \ldots, y_{\pi(t)})$ is simply distributed as $\mathcal{E}_{x_i}^{\mathsf{Enc}}|_t$ (respectively $\mathcal{E}_0^{\mathsf{Enc}}|_t$). Recall that we assume that $U_\pi^n \neq t$, which means that $i \neq n$ or equivalently $x_i = x^*$. Hence, we have

$$\Pr_{\mathbf{y} \sim \mathcal{E}_{\mathbf{x}}^{\mathsf{Enc}}}[\mathcal{A}(\pi(\mathbf{y})) = 1] = \Pr_{\mathbf{y}' \sim \mathcal{E}_{x_i}^{\mathsf{Enc}}|_t}[\mathbf{y}' \in H] = \Pr_{\mathbf{y}' \sim \mathcal{E}_{x^*}^{\mathsf{Enc}}|_t}[\mathbf{y}' \in H].$$

and

$$\Pr_{\mathbf{y} \sim \mathcal{E}_{\mathbf{0}}^{\mathsf{Enc}}}[\mathcal{A}(\pi(\mathbf{y})) = 1] = \Pr_{\mathbf{y}' \sim \mathcal{E}_0^{\mathsf{Enc}}|_t}[\mathbf{y}' \in H].$$

Combining the above two equalities with (13) implies that $\Delta_\pi = \mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|_t, \mathcal{E}_{x^*}^{\mathsf{Enc}}|_t)$ as desired. □

The probability that π falls into this category can be simply computed:

Lemma 8. $\Pr_\pi[C_\pi = t \wedge U_\pi^n \neq t] = \frac{(n-1)\cdot\binom{m}{t}}{\binom{nm}{t}}$.

Proof. $C_\pi = t$ and $|U_\pi^n| \neq t$ if and only if there exists a party $i \in \{1, \ldots, n-1\}$ such that $\pi(\{1, \ldots, t\}) \subseteq \{m(i-1)+1, \ldots, mi\}$. For a fixed i, this happens with probability $\frac{\binom{m}{t}}{\binom{nm}{t}}$. Notice also that the event is disjoint for different i's. As a result, the total probability that this event occurs for at least one i is $(n-1)\cdot\frac{\binom{m}{t}}{\binom{nm}{t}}$. □

Category II: $C_\pi = t$ and $|U_\pi^n| = t$.

We now consider the second category: when $\{\pi(1), \ldots, \pi(t)\}$ corresponds to the last party n. In this case, our choice of x^* implies that $|\Delta_\pi|$ is upper bounded by the statistical distance between $\mathcal{E}_0^{\mathsf{Enc}}|_t$ and $\mathcal{E}_{x^*}^{\mathsf{Enc}}|_t$, as formalized below.

Lemma 9. *For any π such that $C_\pi = t$ and $|U_\pi^n| = t$, we have $|\Delta_\pi| \leq \mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|_t, \mathcal{E}_{x^*}^{\mathsf{Enc}}|_t)$.*

Proof. In this case, we have $\{\pi(1), \cdots, \pi(i)\} \subseteq \{m(n-1)+1, \ldots, mn\}$. Thus, when $\mathbf{y}$ is drawn from $\mathcal{E}_{\mathbf{x}}^{\mathsf{Enc}}$ (respectively $\mathcal{E}_{\mathbf{0}}^{\mathsf{Enc}}$), it is the case that $(y_{\pi(1)}, \ldots, y_{\pi(t)})$ is simply distributed as $\mathcal{E}_{x_n}^{\mathsf{Enc}}|_t$ (respectively $\mathcal{E}_0^{\mathsf{Enc}}|_t$). Hence, we have $\Pr_{\mathbf{y}\sim\mathcal{E}_{\mathbf{x}}^{\mathsf{Enc}}}[\mathcal{A}(\pi(\mathbf{y})) = 1] = \Pr_{\mathbf{y}'\sim\mathcal{E}_{x_n}^{\mathsf{Enc}}|_t}[\mathbf{y}' \in H]$ and $\Pr_{\mathbf{y}\sim\mathcal{E}_{\mathbf{0}}^{\mathsf{Enc}}}[\mathcal{A}(\pi(\mathbf{y})) = 1] = \Pr_{\mathbf{y}'\sim\mathcal{E}_0^{\mathsf{Enc}}|_t}[\mathbf{y}' \in H]$. Combining the above two equalities implies that $|\Delta_\pi| \leq \mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|_t, \mathcal{E}_{x_n}^{\mathsf{Enc}}|_t)$. Recall that x^* is chosen to maximize $\mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|_t, \mathcal{E}_{x^*}^{\mathsf{Enc}}|_t)$, which means that $\mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|_t, \mathcal{E}_{x_n}^{\mathsf{Enc}}|_t)$ $\leq \mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|_t, \mathcal{E}_{x^*}^{\mathsf{Enc}}|_t)$. Hence, we have $|\Delta_\pi| \leq \mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|_t, \mathcal{E}_{x^*}^{\mathsf{Enc}}|_t)$ as desired. □

The probability that π falls into this category can be simply computed in a similar manner as in the first case:

Lemma 10. $\Pr_\pi[C_\pi = t \wedge |U_\pi^n| = t] = \frac{\binom{m}{t}}{\binom{nm}{t}}$.

Proof. $C_\pi = t$ and $|U_\pi^n| = t$ if and only if $\pi(\{1, \ldots, t\}) \subseteq \{m(n-1)+1, \ldots, mn\}$. This happens with probability exactly $\frac{\binom{m}{t}}{\binom{nm}{t}}$. □

Category III: $C_\pi < t$.

Finally, we consider any permutation π such that not all of $\{\pi(1), \ldots, \pi(t)\}$ correspond to a single party. On this front, we may use our choice of t to give an upper bound on $|\Delta_\pi|$ as follows.

Lemma 11. *For any π such that $C_\pi < t$, we have $|\Delta_\pi| < m \cdot \frac{1}{(10nm)^{4(m-C_\pi)}}$.*

Proof. In fact, we will show something even stronger: that the statistical distance of $(y_{\pi(1)}, \ldots, y_{\pi(t)})$ when $\mathbf{y}$ is drawn from $\mathcal{E}_{\mathbf{0}}^{\mathsf{Enc}}$ and that when $\mathbf{y}$ is drawn from $\mathcal{E}_{\mathbf{x}}^{\mathsf{Enc}}$ is at most $m \cdot \frac{1}{(10nm)^{4(m-C_\pi)}}$. The desired bound immediately follows.

Let I denote the set of all parties i such that $U_i \neq \emptyset$. Observe that, when $\mathbf{y}$ is drawn from $\mathcal{E}_{\mathbf{x}}^{\mathsf{Enc}}$ (respectively $\mathcal{E}_{\mathbf{0}}^{\mathsf{Enc}}$), $(y_p)_{p\in U_i}$ is simply distributed as $\mathcal{E}_{x_i}^{\mathsf{Enc}}|_{|U_i|}$ (respectively $\mathcal{E}_0^{\mathsf{Enc}}|_{|U_i|}$) and that these are independent for different i. In other

words, $(y_{\pi(1)}, \ldots, y_{\pi(t)})$ is (after appropriate rearrangement) just the product distribution $\prod_{i\in I} \mathcal{E}_{x_i}^{\mathsf{Enc}}|_{|U_i|}$ (respectively $\prod_{i\in I} \mathcal{E}_{0}^{\mathsf{Enc}}|_{|U_i|}$).

Recall from the definition of C_π that $|U_i|$ is at most C_π for all i. Since $C_\pi < t$ and from our choice of t, we must have $\mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|_{|U_i|}, \mathcal{E}_{x_i}^{\mathsf{Enc}}|_{|U_i|}) < \frac{1}{(10nm)^{4(m-C_\pi)}}$ for all $i \in I$. Hence, we also have

$$SD\left(\prod_{i\in I} \mathcal{E}_0^{\mathsf{Enc}}|_{|U_i|}, \prod_{i\in I} \mathcal{E}_{x_i}^{\mathsf{Enc}}|_{|U_i|}\right) < |I| \cdot \frac{1}{(10nm)^{4(m-C_\pi)}} \le m \cdot \frac{1}{(10nm)^{4(m-C_\pi)}},$$

which concludes the proof. □

Next, we bound the probability that a random permutation π belongs to this category:

Lemma 12. *For all $j < t$, we have $\Pr_\pi[C_\pi = j] \le \frac{n\cdot\binom{m}{t}}{\binom{nm}{t}} \cdot (nm)^{3(t-j)}$.*

Proof. If $C_\pi = j$, there must exist a subset $T \subseteq \{1, \ldots, t\}$ of size j and a party $i \in \{1, \ldots, n\}$ such that $\pi(T) \subseteq \{m(i-1)+1, \ldots, mi\}$. For a fixed T and i, this happens with probability exactly $\frac{\binom{m}{j}}{\binom{nm}{j}}$. Hence, by union bound over all T and i, we have

$$\Pr_\pi[C_\pi = j] \le n \cdot \binom{t}{j} \cdot \frac{\binom{m}{j}}{\binom{nm}{j}} \le \frac{n\cdot\binom{m}{t}}{\binom{nm}{t}} \cdot \frac{\binom{t}{j}\cdot m^{t-j}}{(nm)^{j-t}} \le \frac{n\cdot\binom{m}{t}}{\binom{nm}{t}} \cdot (nm)^{3(t-j)}. \square$$

Putting things together. With all the claims ready, it is now simple to finish the proof of Theorem 7. The difference between the probability that $\mathcal{A}$ accepts on $\mathcal{S}_{\mathbf{0}}^{\mathsf{Enc}}$ and that on $\mathcal{S}_{\mathbf{x}}^{\mathsf{Enc}}$ is

$$\begin{aligned}
\mathop{\mathbb{E}}_\pi[\Delta_\pi] &= \Pr_\pi[C_\pi = t \wedge |U_\pi^n| \neq t] \cdot \mathop{\mathbb{E}}_\pi[\Delta_\pi \mid C_\pi = t \wedge |U_\pi^n| \neq t] \\
&\quad + \Pr_\pi[C_\pi = t \wedge |U_\pi^n| = t] \cdot \mathop{\mathbb{E}}_\pi[\Delta_\pi \mid C_\pi = t \wedge |U_\pi^n| = t] \\
&\quad + \sum_{j=1}^{t-1} \Pr_\pi[C_\pi = j] \cdot \mathop{\mathbb{E}}_\pi[\Delta_\pi \mid C_\pi = j] \\
\text{(Lemmas 7, 8, 9, 10)} &\ge \frac{(n-1)\cdot\binom{m}{t}}{\binom{nm}{t}} \cdot \mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|_t, \mathcal{E}_{x^*}^{\mathsf{Enc}}|_t) - \frac{\binom{m}{t}}{\binom{nm}{t}} \cdot \mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|_t, \mathcal{E}_{x^*}^{\mathsf{Enc}}|_t) \\
&\quad + \sum_{j=1}^{t-1} \Pr_\pi[C_\pi = j] \cdot \mathop{\mathbb{E}}_\pi[\Delta_\pi \mid C_\pi = j] \\
\text{(From } n \ge 3) &\ge \frac{n\cdot\binom{m}{t}}{3\binom{nm}{t}} \cdot \mathrm{SD}(\mathcal{E}_0^{\mathsf{Enc}}|_t, \mathcal{E}_{x^*}^{\mathsf{Enc}}|_t) + \sum_{j=1}^{t-1} \Pr_\pi[C_\pi = j] \cdot \mathop{\mathbb{E}}_\pi[\Delta_\pi \mid C_\pi = j] \\
\text{((13) and Lemma 11)} &\ge \frac{n\cdot\binom{m}{t}}{3\binom{nm}{t}} \cdot \frac{1}{(10nm)^{4(m-t)}} - \sum_{j=1}^{t-1} \frac{\Pr_\pi[C_\pi = j]\cdot m}{(10nm)^{4(m-j)}}
\end{aligned}$$

$$
\begin{aligned}
(\text{Lemma 12}) &\geq \frac{n \cdot \binom{m}{t}}{\binom{nm}{t}} \cdot \left(\frac{1}{3} \cdot \frac{1}{(10nm)^{4(m-t)}} - \sum_{j=1}^{t-1} \frac{(nm)^{3(t-j)} m}{(10nm)^{4(m-j)}} \right) \\
&\geq \frac{n \cdot \binom{m}{t}}{\binom{nm}{t}} \cdot \left(\frac{1}{3} - \sum_{j=1}^{t-1} \frac{1}{10^{t-j}} \right) \cdot \frac{1}{(10nm)^{4(m-t)}} \\
&\geq \frac{n \cdot \binom{m}{t}}{\binom{nm}{t}} \cdot \frac{1}{10} \cdot \frac{1}{(10nm)^{4(m-t)}} \\
&\geq \frac{1}{(nm)^t} \cdot \frac{1}{10} \cdot \frac{1}{(10nm)^{4(m-t)}} \\
&\geq \frac{1}{(10nm)^{5m}}.
\end{aligned}
$$

□

5 Conclusion and Open Questions

In this work, we provide an improved analysis for the split and mix protocol of Ishai et al. [4] in the shuffled model. Our analysis reduces the number of messages required by the protocol by a logarithmic factor. Moreover, for a large range of parameters, we give an asymptotically tight lower bound in terms of the number of messages that each party needs to send for *any* protocol for secure summation.

Although our lower bound is tight in terms of the number of messages, it does not immediately imply any communication lower bound beyond the trivial $\log q$ bound. For instance, when $q = n^{\log n}$ and σ is a constant, then the number of messages needed by Ishai et al.'s protocol is $O\left(\frac{\log q}{\log n}\right) = O(\log n)$ but each message is also of length $O(\log q)$. However, our lower bound does not preclude a protocol with the same number of messages but of length only $O(\log n)$ bits. It remains an interesting open question to close this gap.

Another interesting open question is whether we can give a lower bound for (ε, δ)-differentially private summation protocols when ε is a constant. Currently, our lower bound does not give anything in this regime. In fact, to the best of our knowledge, it remains possible that an $(\varepsilon, 0)$-differentially private summation protocol exists with error $O(1/\varepsilon)$ and where each party sends only $O_\varepsilon(\log n)$ bits. Coming up with such a protocol, or proving that one does not exist, would be a significant step in understanding the power of differential private algorithms in the shuffled model. We point out that following up on this work, [14] studied this question obtaining a pure differentially protocol for summation along with a lower bound, though the tight answer remains unknown.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016

ACM SIGSAC Conference on Computer and Communications Security. pp. 308–318. ACM (2016)
2. Balle, B., Bell, J., Gascón, A., Nissim, K.: Improved summation from shuffling http://arxiv.org/abs/1909.11225
3. Balle, B., Bell, J., Gascón, A., Nissim, K.: Differentially private summation with multi-message shuffling. CoRR **abs/1906.09116** (2019), http://arxiv.org/abs/1906.09116
4. Balle, B., Bell, J., Gascón, A., Nissim, K.: The privacy blanket of the shuffle model. In: Advances in Cryptology - CRYPTO 2019 - 39th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18-22, 2019, Proceedings, Part II. pp. 638–667 (2019)
5. Balle, B., Bell, J., Gascón, A., Nissim, K.: Private summation in the multi-message shuffle model. arXiv: 2002.00817 (2020)
6. Bittau, A., Erlingsson, Ú., Maniatis, P., Mironov, I., Raghunathan, A., Lie, D., Rudominer, M., Kode, U., Tinnés, J., Seefeld, B.: Prochlo: Strong privacy for analytics in the crowd. In: Proceedings of the 26th Symposium on Operating Systems Principles, Shanghai, China, October 28-31, 2017. pp. 441–459 (2017)
7. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for privacy-preserving machine learning. In: Thuraisingham, B.M., Evans, D., Malkin, T., Xu, D. (eds.) Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017. pp. 1175–1191. ACM (2017)
8. Cheu, A., Smith, A.D., Ullman, J., Zeber, D., Zhilyaev, M.: Distributed differential privacy via shuffling. In: Ishai, Y., Rijmen, V. (eds.) Advances in Cryptology - EUROCRYPT 2019 - 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19-23, 2019, Proceedings, Part I. Lecture Notes in Computer Science, vol. 11476, pp. 375–403. Springer (2019)
9. Cormode, G., Garofalakis, M., Haas, P.J., Jermaine, C., et al.: Synopses for massive data: Samples, histograms, wavelets, sketches. Foundations and Trends in Databases **4**(1–3), 1–294 (2011)
10. Corrigan-Gibbs, H., Boneh, D.: Prio: Private, robust, and scalable computation of aggregate statistics. In: NSDI (2017)
11. Erlingsson, Ú., Feldman, V., Mironov, I., Raghunathan, A., Song, S., Talwar, K., Thakurta, A.: Encode, shuffle, analyze privacy revisited: Formalizations and empirical evaluation. arXiv preprint arXiv:2001.03618 (2020)
12. Erlingsson, Ú., Feldman, V., Mironov, I., Raghunathan, A., Talwar, K., Thakurta, A.: Amplification by shuffling: From local to central differential privacy via anonymity. In: SODA. pp. 2468–2479 (2019)
13. Ghazi, B., Golowich, N., Kumar, R., Pagh, R., Velingker, A.: On the power of multiple anonymous messages. arXiv preprint arXiv:1908.11358 (2019)
14. Ghazi, B., Kumar, R., Manurangsi, P., Pagh, R., Velingker, A.: Pure differentially private summation from anonymous messages. arXiv: 2020.01919 (2020)
15. Ghazi, B., Pagh, R., Velingker, A.: Scalable and differentially private distributed aggregation in the shuffled model (2019), http://arxiv.org/abs/1906.08320
16. Goryczka, S., Xiong, L., Sunderam, V.: Secure multiparty aggregation with differential privacy: A comparative study. In: EDBT/ICDT 2013 Workshops
17. Ishai, Y., Kushilevitz, E., Ostrovsky, R., Sahai, A.: Cryptography from anonymity. In: IEEE Symposium on Foundations of Computer Science (FOCS), (2006)

18. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R.G.L., Rouayheb, S.E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P.B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S.U., Sun, Z., Suresh, A.T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F.X., Yu, H., Zhao, S.: Advances and open problems in federated learning. arXiv: 1912.04977 (2019)
19. Kearns, M.: Efficient noise-tolerant learning from statistical queries. JACM (1998)
20. Kenthapadi, K., Korolova, A., Mironov, I., Mishra, N.: Privacy via the johnson-lindenstrauss transform. Journal of Privacy and Confidentiality **5** (2013)
21. Konečnỳ, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., Bacon, D.: Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492 (2016)
22. Liu, A., Xia, L., Duchowski, A., Bailey, R., Holmqvist, K., Jain, E.: Differential privacy for eye-tracking data. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications. pp. 1–10 (2019)
23. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. pp. 1273–1282 (2017)
24. McMahan, H.B., Ramage, D.: Federated learning: Collaborative machine learning without centralized training data. Google AI Blog (April 2017), https://ai.googleblog.com/2017/04/federated-learning-collaborative.html
25. McMahan, H.B., Ramage, D., Talwar, K., Zhang, L.: Learning differentially private recurrent language models. arXiv preprint arXiv:1710.06963 (2017)
26. Melis, L., Danezis, G., Cristofaro, E.D.: Efficient private statistics with succinct sketches. In: NDSS (2016)
27. Mishra, N., Sandler, M.: Privacy via pseudorandom sketches. In: PODS (2006)
28. Reyzin, L., Smith, A.D., Yakoubov, S.: Turning hate into love: Homomorphic ad hoc threshold encryption for scalable mpc. IACR Cryptology ePrint Archive **2018**
29. Steil, J., Hagestedt, I., Huang, M.X., Bulling, A.: Privacy-aware eye tracking using differential privacy. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications. pp. 1–9 (2019)
30. Wang, T., Xu, M., Ding, B., Zhou, J., Li, N., Jha, S.: Practical and robust privacy amplification with multi-party differential privacy. arXiv:1908.11515 (2019)
31. Woodruff, D.P., et al.: Sketching as a tool for numerical linear algebra. Foundations and Trends in Theoretical Computer Science **10**(1–2), 1–157 (2014)

A Proofs of Bounds for Multinomial Coefficients

Below we prove Facts 1 and 2 from Section 3.

Proof of Fact 1. Let $U = [a_1 + a'_1 + \cdots + a_k + a'_k], A = [a_1 + \cdots + a_k]$ and $B = U \setminus A$.

Consider the following process of generating a partition $S_1 \sqcup \cdots \sqcup S_k = U$. First, take a partition $T_1 \sqcup \cdots \sqcup T_k = A$ and a partition $T'_1 \sqcup \cdots \sqcup T'_k = B$. Then, let $S_i = T_i \cup T'_i$ for all $i \in [k]$.

Notice that each pair of $T_1 \sqcup \cdots \sqcup T_k$ with $|T_i| = a_i$ and $T'_1 \sqcup \cdots \sqcup T'_k$ with $|P_i| = a'_i$ produces different $S_1 \sqcup \cdots \sqcup S_k = U$ with $|S_i| = a_i + a'_i$. Since the number of such pairs $T_1 \sqcup \cdots \sqcup T_k$ and $T'_1 \sqcup \cdots \sqcup T'_k$ is $\binom{a_1+\cdots+a_k}{a_1,\ldots,a_k} \cdot \binom{a'_1+\cdots+a'_k}{a'_1,\ldots,a'_k}$ and the number of $S_1 \sqcup \cdots \sqcup S_k = U$ with $|S_i| = a_i + a'_i$ is only $\binom{a_1+\cdots+a_k+a'_1+\cdots+a'_k}{a_1+a'_1,\ldots,a_k+a'_k}$, we have

$$\binom{a_1 + \cdots + a_k + a'_1 + \cdots + a'_k}{a_1 + a'_1, \ldots, a_k + a'_k} \geq \binom{a_1 + \cdots + a_k}{a_1, \ldots, a_k} \cdot \binom{a'_1 + \cdots + a'_k}{a'_1, \ldots, a'_k}$$

as desired. □

Proof of Fact 2. Assume w.l.o.g. that $a_1 \leq a_2 \leq \cdots \leq a_k$. We have

$$\begin{aligned}
\binom{a_1 + \cdots + a_k}{a_1, \ldots, a_k} = \prod_{i=1}^{k} \binom{a_i + \cdots + a_k}{a_i} &\geq \prod_{i=1}^{\lfloor k/2 \rfloor} \binom{a_i + \cdots + a_k}{a_i} \\
&\geq \prod_{i=1}^{\lfloor k/2 \rfloor} (a_i + \cdots + a_k) \\
&\geq \left(\frac{a_1 + \cdots + a_k}{2} \right)^{\lfloor k/2 \rfloor},
\end{aligned}$$

where the last inequality uses the fact that $a_1 \leq \cdots \leq a_k$. □

B Proof of Corollary 1

Corollary 1 follows from our main theorem (Theorem 1) and the connection between secure summation protocols and differentially private summation protocols due to Balle et al. [3]. We recall the latter below.

Lemma 13 (Lemma 4.1 of [3]). *Given a σ-secure protocol in the anonymized setting for n-party summation over the domain $\mathbb{F}_q$, where each party sends $f(q, n, \sigma)$ messages each of $g(q, n, \sigma)$ bits, there exists an $(\varepsilon, (1 + e^{\varepsilon})2^{-\sigma-1})$-differentially private protocol in the shuffled model for real summation with absolute error $O(1 + 1/\varepsilon)$ where each party sends $f(O(n^{3/2}), n, \sigma)$ messages each of $g(O(n^{3/2}), n, \sigma)$ bits.*

Corollary 1 now follows immediately by applying Lemma 13 and Theorem 1 with $\sigma = 1 + \log\left(\frac{1+e^{\varepsilon}}{\delta}\right) = O\left(1 + \varepsilon + \log(1/\delta)\right)$.

We remark here that Lemma 13 as stated above is slightly different from Lemma 4.1 of [3]. In particular, in [3], the statement requires the secure summation protocol to works for any $\mathbb{Z}_q$ even when q is not a prime power. On the other hand, our analysis in this paper (which uses rank of matrices) only applies to when q is a prime power (i.e., $\mathbb{F}_q$ is a field). However, it turns out that this does not affect the connection too much: instead of picking $q = 2\lceil n^{3/2} \rceil$ as in [3], we may pick q to be the smallest prime larger than $2n^{3/2}$. In this case, q remains $O(n^{3/2})$ and the remaining argument of [3] remains exactly the same.