

# Differentially Private Aggregation via Imperfect Shuffling

**Badih Ghazi** ✉

Google Research, Mountain View, CA, USA

**Ravi Kumar** ✉

Google Research, Mountain View, CA, USA

**Pasin Manurangsi** ✉

Google Research, Mountain View, CA, USA

**Jelani Nelson** ✉

University of California at Berkeley, CA, USA

Google Research, Mountain View, CA, USA

**Samson Zhou** ✉

University of California at Berkeley, CA, USA

Rice University, Houston, TX, USA

---

## Abstract

In this paper, we introduce the imperfect shuffle differential privacy model, where messages sent from users are shuffled in an *almost* uniform manner before being observed by a curator for private aggregation. We then consider the private summation problem. We show that the standard split-and-mix protocol by Ishai et. al. [FOCS 2006] can be adapted to achieve near-optimal utility bounds in the imperfect shuffle model. Specifically, we show that surprisingly, there is no additional error overhead necessary in the imperfect shuffle model.

**2012 ACM Subject Classification** Security and privacy → Human and societal aspects of security and privacy

**Keywords and phrases** Differential privacy, private summation, shuffle model

**Digital Object Identifier** 10.4230/LIPIcs.ITC.2023.17

## 1 Introduction

Differential privacy (DP) [18] has emerged as a popular concept that mathematically quantifies the privacy of statistics-releasing mechanisms. Consequently, DP mechanisms have been recently deployed in industry [29, 21, 35, 16], as well as by government agencies such as the US Census Bureau [2]. DP is parameterized by  $\epsilon$  and  $\delta$ , where  $\epsilon$  is a privacy loss parameter that is generally a small positive constant such as 1 and  $\delta$  is an approximation parameter or “failure” probability that is typically (smaller than) inverse-polynomial in  $n$ :

► **Definition 1** (Differential privacy; [18, 17]). *Given  $\epsilon > 0$  and  $\delta \in (0, 1)$ , a randomized algorithm  $\mathcal{A} : X \rightarrow Y$  is  $(\epsilon, \delta)$ -differentially private if, for every neighboring datasets  $x$  and  $x'$ , and for all  $S \subseteq Y$ ,  $\Pr[\mathcal{A}(x) \in S] \leq e^\epsilon \cdot \Pr[\mathcal{A}(x') \in S] + \delta$ .*

In this paper, we study the real summation problem, where each of  $n$  parties holds a number  $x_i \in [0, 1]$  for all  $i \in [n]$  and the goal is to privately (approximately) compute  $\sum_{i=1}^n x_i$ . Due to its fundamental nature, the private real summation problem has a wide range of applications, such as private distributed mean estimation [40, 10], e.g., in federated learning [33, 27, 31], private stochastic gradient descent [37, 8, 1, 3, 13], databases and information systems [34, 43], and clustering [39, 38].



© Badih Ghazi, Ravi Kumar, Pasin Manurangsi, Jelani Nelson, and Samson Zhou;  
licensed under Creative Commons License CC-BY 4.0

4th Conference on Information-Theoretic Cryptography (ITC 2023).

Editor: Kai-Min Chung; Article No. 17; pp. 17:1–17:22



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In the central model of DP, where a curator is given full access to the raw data in order to release the private statistic or data structure, the Laplace mechanism [18] can achieve, for real summation, additive error  $\mathcal{O}(\frac{1}{\epsilon})$ , which is known to be nearly optimal for  $\epsilon \leq 1$  [26].

However, the ability for the curator to observe the full data is undesirable in many commercial settings, where the users do not want their raw data to be sent to a central curator. To address this shortcoming, the local model of DP [32, 42] (LDP) demands that all messages sent from an individual user to the curator is private. Unfortunately, although the local model enjoys near-minimal trust assumptions, numerous basic tasks provably must suffer from significantly larger estimation errors compared to their counterparts in the central model. For the real summation problem, [9] achieves additive error  $\mathcal{O}_\epsilon(\sqrt{n})$  and it is known that smaller error bounds cannot be achieved [12].

Consequently, the shuffle model [11, 20, 14] of DP was introduced as an intermediary between the generous central model and the strict local model. In the shuffle model, the messages sent from the users are randomly permuted before being observed by the curator, in an encode-shuffle-analyze architecture. Surprisingly, when users are allowed to send multiple messages, there exist protocols in the shuffle model of DP that achieve additive error  $\mathcal{O}_\epsilon(1)$  for the private real summation problem [25, 6, 24]. Unfortunately, practical applications can lack the ideal settings that provide the full assumptions required by the shuffle model of DP.

## 1.1 Model and Motivation

We first define a natural generalization of the uniform shuffler that tolerates imperfections. Let  $\Pi$  be the set of permutations on  $[n]$ . For  $\pi, \pi' \in \Pi$ , we define  $\text{Swap}(\pi, \pi')$  to be the minimum number of coordinate *swaps*<sup>1</sup> that can be applied to  $\pi$  to obtain  $\pi'$ .

► **Definition 2** ( $\gamma$ -Imperfect Shuffler). *For a distortion parameter  $\gamma > 0$ , we say that  $\mathcal{S}$  is a  $\gamma$ -imperfect shuffler if, for all  $\pi, \pi' \in \Pi$ ,*

$$\Pr[\mathcal{S} = \pi] \leq e^{\gamma \cdot \text{Swap}(\pi, \pi')} \Pr[\mathcal{S} = \pi'].$$

We call an output from such a shuffler a  $\gamma$ -imperfect shuffle or a  $\gamma$ -I-shuffle, for short. Here,  $\gamma$  represents an upper bound on the multiplicative distortion of the output probabilities of the distributions of the shuffler, i.e., how the distribution deviates from a perfectly symmetric shuffler. For example,  $\gamma = 0$  corresponds to a perfectly symmetric shuffler while  $\gamma \rightarrow \infty$  represents almost no guarantee from the shuffler.

To understand the motivation behind this definition, consider a setting where a number of user devices collect statistics to be sent to an intermediate buffer, which is ultimately sent to a central curator for processing. The devices may choose to perform this collection over different periods of time, so that immediately sending their statistics over to the curator could reveal information about their identity, through the timestamp.

For example, consider a setting where sensors are monitoring traffic in US cities during peak afternoon hours. Then reports that are received earlier in the day by the curator are more likely to correspond to cities that are in the east, while reports that are received later in the day by the curator are more likely to correspond to cities in the west. To mitigate this, the sensors instead could choose a universally fixed hour during the day to broadcast their reports from the previous day, at some random time during the hour.

<sup>1</sup> We say that  $\pi'$  results from an application of a coordinate *swap* on  $\pi$  iff  $\pi(i) = \pi'(i)$  on all except two  $i \in [n]$ .

Specifically, each user  $i \in [n]$  could choose a time  $t_i$ , say normalized without loss of generality to  $t_i \in [0, 1]$ , and send their messages at time  $t_i$ . If the  $t_i$  are chosen uniformly at random and this protocol was executed perfectly, it would result in a uniform shuffle of the messages for a buffer that strips both the source information and the exact time of arrival, e.g., [41].<sup>2</sup> However, issues may arise such as different clock skews, where users may not perfectly synchronize the fixed hour during which the messages should be sent, or communication delays, either because an intermediate link has failed or simply because the latency varies across different networks. It is unclear how to model the imperfect shuffle resulting from these issues using the standard shuffle model.

For a better handle on modeling the imperfection, we can assume that each  $t_i$  is adversarially chosen in  $[0, 1]$ . Moreover, each message transmission time can now be altered by a random offset from the intended release time, where the offset is drawn, e.g., from a Laplacian distribution. Specifically, each user  $i \in [n]$  draws an offset  $\tau_i$  from the (centered) Laplacian distribution with scale  $\frac{2}{\gamma}$ , and sends their message at time  $t_i + \tau_i$  instead of at time  $t_i$ .

In other words, each user  $i \in [n]$  sends their message at time  $t_i + \tau_i$ , which is determined by the two following quantities:

- (1)  $t_i$  is an arbitrary and possibly adversarially chosen offset due to nature or some other external source, e.g., clock skews, transmission failure, communication delay.
- (2)  $\tau_i$  is an internal source of noise that the protocol can sample from a predetermined distribution to mitigate the negative privacy effects of  $t_i$ .

Note that whereas two permutations  $\pi, \pi'$  on  $[n]$  with swap distance one were equally likely to be output by the shuffler, this may now no longer be the case. On the other hand, for fixed  $i, j \in [n]$  and conditioning on the values of  $\{t_1, \tau_1, \dots, t_n, \tau_n\} \setminus \{t_i, \tau_i, t_j, \tau_j\}$ , we can see that for  $a, b \in [n]$ , the probability that  $t_a + \tau_a \leq t_i + \tau_i \leq t_{a+1} + \tau_{a+1}$  and  $t_b + \tau_b \leq t_j + \tau_j \leq t_{b+1} + \tau_{b+1}$  is within an  $e^\gamma$  factor of the probability that  $t_a + \tau_a \leq t_j + \tau_j \leq t_{a+1} + \tau_{a+1}$  and  $t_b + \tau_b \leq t_i + \tau_i \leq t_{b+1} + \tau_{b+1}$ .

Specifically, let  $\mathcal{E}_1$  be the event that  $\tau_i \in [t_a + \tau_a - t_i, t_{a+1} + \tau_{a+1} - t_i]$ , where  $\tau_i$  is a (centered) Laplace random variable and scale  $\frac{2}{\gamma}$ . Similarly, let  $\mathcal{E}_2$  be the event that  $\tau_j \in [t_b + \tau_b - t_j, t_{b+1} + \tau_{b+1} - t_j]$  where  $\tau_j$  is a (centered) Laplace random variable and scale  $\frac{2}{\gamma}$ . Furthermore, let  $\mathcal{E}_3$  be the event that  $\tau_j \in [t_a + \tau_a - t_j, t_{a+1} + \tau_{a+1} - t_j]$  and  $\mathcal{E}_4$  be the event that  $\tau_i \in [t_b + \tau_b - t_i, t_{b+1} + \tau_{b+1} - t_i]$ . Then by the properties of the Laplace distribution and the assumption that  $t_i, t_j \in [0, 1]$ , we have  $\Pr[\mathcal{E}_1 \wedge \mathcal{E}_2] = \Pr[\mathcal{E}_1] \Pr[\mathcal{E}_2] \leq (e^{\gamma/2} \cdot \Pr[\mathcal{E}_3])(e^{\gamma/2} \cdot \Pr[\mathcal{E}_4]) = e^\gamma \cdot \Pr[\mathcal{E}_3 \wedge \mathcal{E}_4]$ . Thus, the resulting distribution over permutations is captured by the  $\gamma$ -I-shuffle model.

We can naturally generalize this setting to the model where each user sends  $m$  messages, e.g.,  $m$  buffers collect messages from  $n$  users, which results in times  $\{t_{i,j}\}_{i \in [n], j \in [m]}$  and offsets  $\{\tau_{i,j}\}_{i \in [n], j \in [m]}$ . Formally, for  $m$  rounds of messages for the  $n$  users,  $\{m_{i,j}\}_{i \in [n], j \in [m]}$ , a separate permutation  $\pi_j$  drawn from a  $\gamma$ -imperfect shuffler is used to shuffle the messages  $\{m_{i,j}\}_{i \in [n]}$ , for each  $j \in [m]$ . For example,  $\{m_{i,1}\}_{i \in [n]}$  is shuffled according to a permutation  $\pi_1$  drawn from a  $\gamma$ -imperfect shuffler,  $\{m_{i,2}\}_{i \in [n]}$  is shuffled according to an independent permutation  $\pi_2$  drawn from the same  $\gamma$ -imperfect shuffler, and so on and so forth.

<sup>2</sup> We assume in this example that the buffer can queue the messages, and then forward them to the analyst at some point of time, but that it cannot further shuffle them. The (imperfect) shuffling we consider stems solely from the randomization of the transmission time of the messages by the users.

We remark that the above model is sometimes referred to as the *m-parallel shuffling* model; another model that has been considered in literature is one where all the  $mn$  messages are shuffled together using a single shuffler. We only focus on the former in this paper. It remains an interesting open question whether our results can be extended to the latter model.

## 1.2 Our Contributions

Surprisingly, we present a protocol for the real summation problem that matches the utility bounds of the best protocols in the shuffle model. Thus, we show that there is no additional error overhead necessary in the  $\gamma$ -I-shuffle model, i.e., there is no utility loss due to the imperfect shuffler.

► **Theorem 3.** *Let  $n \geq 19$  and  $\gamma \leq \frac{\log \log n}{80}$  be a distortion parameter. Then there exists an  $(\varepsilon, \delta)$ -DP protocol for summation in the  $\gamma$ -I-shuffle model with expected absolute error  $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$  and  $m = \mathcal{O}\left(e^{4\gamma} + \frac{e^{4\gamma}(\log \frac{1}{\varepsilon} + \log n)}{\log n}\right)$  messages per party. Each message uses  $\mathcal{O}(\log q)$  bits, for  $q = \lceil 2n^{3/2} \rceil$ .*

Observe that when  $\delta$  is inverse-polynomial in  $n$  and the distortion parameter  $\gamma$  is a constant  $\mathcal{O}(1)$ , then the number of messages  $m$  sent by each player in Theorem 3 is a constant. Moreover, under these settings, Theorem 3 recovers the guarantees in the standard shuffle model from [6, 25], though we remark that more communication efficient protocols [24] are known in the standard shuffle model across more general settings. Regardless, we again emphasize that the privacy and utility guarantees of the protocol are independent of the distortion parameter  $\gamma$ .

## 1.3 Overview of our Techniques

In this section, we describe both our protocol for private real summation in the  $\gamma$ -I-shuffle model and the corresponding analysis for correctness and privacy.

A natural starting point is the recent framework by [44, 45], which achieves amplification of privacy using *differentially oblivious* (DO) shufflers that nearly match amplification of privacy results using fully anonymous shufflers [20, 4, 14, 22]. Unfortunately, the framework crucially uses LDP protocols, which are known to not give optimal bounds even with fully anonymous shufflers. For instance, [6, 14, 5] showed that any single-message shuffled protocol for summation based on LDP protocols must exhibit mean squared error  $\Omega(n^{1/3})$  or absolute error  $\Omega(n^{1/6})$ .

Another natural approach is to adapt recent works for private real summation in the shuffle model, e.g., [25, 24]. One challenge in generalizing these proofs is that they often leverage the fully anonymous shuffler by analyzing a random sample from an alternate view of the output of the local randomizers, which often have some algebraic or combinatorial interpretation that facilitates the proof of specific desirable properties. However, these properties often seem substantially more difficult to prove once the symmetry of the fully anonymous shuffler is lost. In fact, we do not even know the mass that the  $\gamma$ -imperfect shuffler places on each permutation.

**From private real summation to statistical security of summation on fixed fields.** We first use an observation from [6] that reduces the problem of private real summation to the problem of private summation on a fixed field of size  $q$ , so that each user has an input  $x_i \in \mathbb{F}_q$  for all  $i \in [n]$ . We then consider the well-known split-and-mix protocol [30], where

each user  $i$  outputs a set of  $m$  messages  $x_{i,1}, \dots, x_{i,m} \in \mathbb{F}_q$  uniformly at random conditioned on  $x_{i,1} + \dots + x_{i,m} = x_i$ . For the private summation on a fixed field problem, we adapt a well-known reduction [6] for the split-and-mix protocol in the shuffle DP model to the notion of statistical security in the  $\gamma$ -I-shuffle model. Statistical security demands small total variation between the output of a protocol on input  $x$  and input  $x'$ , if  $\sum_{i=1}^n x_i = \sum_{i=1}^n x'_i$ . In other words, it suffices to show that the output distribution looks “similar” for two inputs with the same sum. See Definition 5 for a formal definition of statistical security.

To show statistical security, we first upper-bound the total variation distance in terms of the probability that two independent instances of the same protocol with the *same* input give the same output. Balle et al. [6] use a similar approach, but then utilizes the symmetry of the fully anonymous shuffler to further upper-bound this quantity in terms of the probability that  $\vec{\mathcal{R}}(\vec{X}) = \mathcal{S} \circ \vec{\mathcal{R}}'(\vec{X})$ , where  $\vec{X} = (x_1, \dots, x_n)$  is the input vector,  $\vec{\mathcal{R}}$  and  $\vec{\mathcal{R}}'$  are independent instances of the local randomizer, and  $\mathcal{S}$  is an instance of the uniform shuffler. We do not have access to such symmetries in the  $\gamma$ -I-shuffle model or even explicit probabilities that the  $\gamma$ -imperfect shuffler places on each permutation.

**Connected components of a communication graph.** Instead, we first upper-bound the total variation distance by  $\vec{\mathcal{R}}(\vec{X}) = \mathcal{S}^{-1} \circ \mathcal{S}' \circ \vec{\mathcal{R}}'(\vec{X})$ , where  $\mathcal{S}^{-1}$  is the inverse of an instance of a  $\gamma$ -imperfect shuffle and  $\mathcal{S}'$  is an independent instance of the same  $\gamma$ -imperfect shuffle. Intuitively,  $\vec{\mathcal{R}}(\vec{X})$  and  $\mathcal{S}^{-1} \circ \mathcal{S}' \circ \vec{\mathcal{R}}'(\vec{X})$  can look very different if there exists a large number of users whose messages are not shuffled with those of other users. Formally, this can be captured by looking at the number of connected components in the communication graph of  $\mathcal{S}^{-1} \circ \mathcal{S}' \circ \vec{\mathcal{R}}'(\vec{X})$ , so that there exists an edge connecting users  $i$  and  $j$  if the protocol swaps one of their messages. Hence, evaluating the number of connected components in the communication graph is closely related to analyzing the probability that there is no edge between  $S$  and  $[n] \setminus S$ , for a given set  $S \subseteq [n]$ .

Although this quantity would be somewhat straightforward to evaluate for a uniform shuffler [6], it seems more challenging to evaluate for  $\gamma$ -imperfect shufflers, since we do not have explicit probabilities for each permutation. Therefore, we develop a novel coupling argument to relate the probability that there is no edge between  $S$  and  $[n] \setminus S$  in the  $\gamma$ -I-shuffle model to the probability of this event in the shuffle model. In particular, a specific technical challenge that our argument handles is when both  $S$  and  $[n] \setminus S$  has large cardinality, because then there can be a permutation  $\pi$  that swaps many coordinates while still leaving  $S$  and  $[n] \setminus S$  disconnected. However, if we simply relate the probability of  $\Pi$  in the shuffle and the  $\gamma$ -I-shuffle model, we incur a gap of  $e^{t \cdot \gamma}$ , where  $\gamma$  is the distortion parameter and  $t$  is the number of swaps by  $\Pi$ , which can have size  $\Omega(n)$ . Thus without additional care, this gap can overwhelm the probability achieved from the coupling argument. We circumvent this issue by considering a subset of  $S$  with size  $k$  and coupling the “good” permutations in the shuffle and the  $\gamma$ -I-shuffle model, which results in a smaller gap of  $e^{k \cdot \gamma}$ . For more details, see Lemma 26.

**Putting things together.** At this point, we are almost done. Unfortunately, our coupling only addresses the case where a single imperfect shuffle is performed on a local randomizer, but we require the bound for the composition  $\mathcal{S}^{-1} \circ \mathcal{S}' \circ \vec{\mathcal{R}}'(\vec{X})$ , which seems significantly more challenging because communication between users  $i$  and  $j$  under  $\mathcal{S}'$  may be “erased” by  $\mathcal{S}^{-1}$ . Instead, we show a simple observation for  $\gamma$ -imperfect shuffling, which states that if  $\mathcal{S}, \mathcal{S}'$  are two shufflers such that  $\mathcal{S}$  is a  $\gamma$ -imperfect shuffler, then  $\mathcal{S}' \circ \mathcal{S}$  is a  $\gamma$ -imperfect shuffler. This statement, presented in Lemma 23, can be considered as a post-processing preservation

property of  $\gamma$ -imperfect shuffling. In light of this statement, we can now view  $\mathcal{S}^{-1} \circ \mathcal{S}' \circ \vec{\mathcal{R}}'(\vec{X})$  as a single  $\gamma$ -imperfect shuffler applied to the local randomizer, and use our new results upper-bounding the number of connected components in the resulting communication graph to ultimately show  $\sigma$ -security.

## 1.4 Preliminaries

For an integer  $n > 0$ , we define  $[n] := \{1, \dots, n\}$ .

► **Definition 4** (Total variation distance). *Given probability measures  $\mu, \nu$  on a domain  $\Omega$ , their total variation distance is defined by  $\text{TVD}(\mu, \nu) = \frac{1}{2} \|\mu - \nu\|_1 = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|$ .*

► **Definition 5** ( $\sigma$ -security). *Given a security parameter  $\sigma > 0$ , a protocol  $\mathcal{P}$  is  $\sigma$ -secure for computing a function  $f : \mathcal{X}^n \rightarrow Z$  if, for any  $x, x' \in \mathcal{X}^n$  such that  $f(x) = f(x')$ , we have  $\text{TVD}(\mathcal{P}(x), \mathcal{P}(x')) \leq 2^{-\sigma}$ .*

Recall the following two well-known properties of DP:

► **Theorem 6** (Basic Composition, e.g., [19]). *Let  $\varepsilon, \delta \geq 0$ . Any mechanism that permits  $k$  adaptive interactions with mechanisms that preserve  $(\varepsilon, \delta)$ -DP is  $(k\varepsilon, k\delta)$ -DP.*

► **Theorem 7** (Post-processing [19]). *Let  $\mathcal{M} : \mathcal{U}^* \rightarrow X$  be an  $(\varepsilon, \delta)$ -DP algorithm. Then, for any arbitrary random mapping  $g : X \rightarrow X'$ , we have that  $g(\mathcal{M}(x))$  is  $(\varepsilon, \delta)$ -DP.*

We use  $\text{Ber}(p)$  to denote the Bernoulli distribution with parameter  $p$  and use  $\text{DLap}(\alpha)$  to denote the discrete Laplace distribution, so that  $Z \sim \text{DLap}(\alpha)$  has the probability mass function  $\Pr[Z = k] \propto \alpha^{|k|}$  for  $k \in \mathbb{Z}$ . We use  $\text{Polya}(r, p)$  to denote the Polya distribution with parameter  $r > 0, p \in (0, 1)$ , which induces the probability density function  $k \mapsto \binom{k+r-1}{k} p^k (1-p)^r$  for  $k \in \mathbb{Z}_{\geq 0}$ . We require the following equivalence between a discrete Laplacian random variable and the sum of a differences of Polya random variables.

► **Fact 8.** *Let  $x_1, \dots, x_n, y_1, \dots, y_n \sim \text{Polya}(\frac{1}{n}, \alpha)$ . Then  $z = \sum_{i=1}^n (x_i - y_i) \sim \text{DLap}(\alpha)$ .*

We also require the following property about randomized rounding.

► **Lemma 9** ([4]). *Given a precision  $p \geq 1$ , let  $x_1, \dots, x_n \in \mathbb{R}$  and  $y_i = \lfloor x_i p \rfloor + \text{Ber}(x_i p - \lfloor x_i p \rfloor)$  for each  $i \in [n]$ . Then  $\mathbb{E} \left[ \left( \sum_{i=1}^n \left( x_i - \frac{y_i}{p} \right) \right)^2 \right] \leq \frac{n}{4p^2}$ .*

## 1.5 Related Work

To amplify the privacy, the trusted shuffler is the key component of the shuffle model, which in some sense only shifts the point of vulnerability from the curator to the shuffler, particularly in the case where the shuffler may be colluding with the curator. Hence among the various relaxations for distributed DP protocols, e.g., [7, 15], the DO shuffle model has been recently proposed [36, 28] to permit some privacy leakage in the shuffling stage, called a DO shuffle. In fact, [36, 28] showed that DO shuffling can be more efficient to achieve than a fully anonymous shuffle while [44, 45] showed that locally private protocols can be used in conjunction with a DO shuffler to achieve almost the same privacy amplification bounds as with a fully anonymous shuffler, up to a small additive loss resulting from the DO shuffle. However, the best known results in the shuffle model of DP do not utilize LDP protocols, and thus cannot directly be applied in the framework of [44, 45].

## 2 A Simple Reduction

In this section, we briefly describe a simple reduction for showing amplification of privacy for imperfect shuffling. The result can be viewed as in the same spirit as similar privacy amplification statements, e.g., [22, 44, 23], but for imperfect shuffling. In particular, the following well-known result achieves privacy amplification for local randomizers in the shuffle model:

► **Theorem 10** ([22]). *For any domain  $\mathcal{D}$  and  $i \in [n]$ , let  $\mathcal{R}^{(i)} : \mathcal{S}^{(1)} \times \dots \times \mathcal{S}^{(i-1)} \times \mathcal{D} \rightarrow \mathcal{S}^{(i)}$ , where  $\mathcal{S}^{(i)}$  is the range space of  $\mathcal{R}^{(i)}$ , such that  $\mathcal{R}^{(i)}(z_{1:i-1}, \cdot)$  is an  $\varepsilon_0$ -DP local randomizer for all values of auxiliary inputs  $z_{1:i-1} \in \mathcal{S}^{(1)} \times \dots \times \mathcal{S}^{(i-1)}$ . Let  $\mathcal{A}_s : \mathcal{D}^n \rightarrow \mathcal{S}^{(1)} \times \dots \times \mathcal{S}^{(n)}$  be the algorithm that given a dataset  $x_{1:n} \in \mathcal{D}^n$ , samples a uniform random permutation  $\pi$  over  $[n]$  and sequentially computes  $z_i = \mathcal{R}^{(i)}(z_{1:i-1}, x_{\pi(i)})$  for  $i \in [n]$  and outputs  $z_{1:n}$ . Then for any  $\delta \in [0, 1]$  such that  $\varepsilon_0 \leq \log\left(\frac{n}{16 \log(2/\delta)}\right)$ ,  $\mathcal{A}_s$  is  $(\varepsilon, \delta)$ -DP for  $\varepsilon \leq \log\left(1 + \frac{e^{\varepsilon_0} - 1}{e^{\varepsilon_0} + 1} \left(\frac{8\sqrt{e^{\varepsilon_0} \log(4/\delta)}}{\sqrt{n}}\right)\right)$ .*

We would like to show privacy amplification statements for the imperfect shuffle model that are qualitatively similar to Theorem 10. To that end, we first recall the following definition of DO shufflers.

► **Definition 11** (Differentially Oblivious Shuffle). *A shuffle protocol is  $(\varepsilon, \delta)$ -differentially oblivious if for all adversaries  $\mathcal{V}$ , all  $\pi, \pi' \in \Pi$ , and all subsets  $S$  of the view space,  $\Pr[\text{View}^{\mathcal{V}}(\pi) \in S] \leq e^{\varepsilon \cdot \text{Swap}(\pi, \pi')} \Pr[\text{View}^{\mathcal{V}}(\pi') \in S] + \delta$ .*

Zhou and Shi [44] showed that DO shufflers also amplify privacy.

► **Theorem 12** (Theorem 1 in [44]). *For any domain  $\mathcal{D}$  and range space  $\mathcal{S}$ ,  $i \in [n]$ , let  $\mathcal{R}^{(1)}, \dots, \mathcal{R}^{(n)} : \mathcal{D} \rightarrow \mathcal{S}$  be  $\varepsilon_0$ -DP local randomizers and let  $\mathcal{A}_s$  be a  $(\varepsilon_1, \delta_1)$ -DO shuffler. Then the composed protocol  $\mathcal{A}_s(\mathcal{R}^{(1)}, \dots, \mathcal{R}^{(n)})$  is  $(\varepsilon + \varepsilon_1, \delta + \delta_1)$ -DP for  $\varepsilon = \mathcal{O}\left(\frac{(1 - e^{\varepsilon_0})e^{\varepsilon_0/2}\sqrt{\log(1/\delta)}}{\sqrt{n}}\right)$ .*

It turns out that imperfect shufflers can be parametrized by DO shufflers, i.e., imperfect shufflers are a specific form of DO shufflers. Therefore, we can immediately apply the previous statement to obtain the following statement for privacy amplification for imperfect shufflers.

► **Theorem 13.** *For any domain  $\mathcal{D}$  and range space  $\mathcal{S}$ ,  $i \in [n]$ , let  $\mathcal{R}^{(1)}, \dots, \mathcal{R}^{(n)} : \mathcal{D} \rightarrow \mathcal{S}$  be  $\varepsilon_0$ -DP local randomizers and let  $\mathcal{A}_s$  be a  $\gamma$ -imperfect shuffler. Then the composed protocol  $\mathcal{A}_s(\mathcal{R}^{(1)}, \dots, \mathcal{R}^{(n)})$  is  $(\varepsilon + \gamma, \delta)$ -DP for  $\varepsilon = \mathcal{O}\left(\frac{(1 - e^{\varepsilon_0})e^{\varepsilon_0/2}\sqrt{\log(1/\delta)}}{\sqrt{n}}\right)$ .*

**Proof.** By the definition of  $\gamma$ -imperfect shuffle, we have that for all  $\pi, \pi' \in \Pi$ ,  $\Pr[S = \pi] \leq e^{\gamma \cdot \text{Swap}(\pi, \pi')} \Pr[S = \pi']$ . Since no additional information is leaked by the shuffler, then for all adversaries  $\mathcal{V}$  and all subsets  $S$  of the view space,

$$\Pr[\text{View}^{\mathcal{V}}(\pi) \in S] \leq e^{\gamma \cdot \text{Swap}(\pi, \pi')} \Pr[\text{View}^{\mathcal{V}}(\pi') \in S].$$

In other words, the  $\gamma$ -imperfect shuffler is a  $(\gamma, 0)$ -DO shuffler. Thus by Theorem 12, the composed protocol  $\mathcal{A}_s(\mathcal{R}^{(1)}, \dots, \mathcal{R}^{(n)})$  is  $(\varepsilon + \gamma, \delta)$ -DP for  $\varepsilon = \mathcal{O}\left(\frac{(1 - e^{\varepsilon_0})e^{\varepsilon_0/2}\sqrt{\log(1/\delta)}}{\sqrt{n}}\right)$ . ◀



### 3 Differentially Private Summation

In this section, we first introduce the structural statements necessary to argue privacy for the standard split-and-mix protocol [30]. We then assume correctness of these statements, deferring their proofs to subsequent sections, and we prove the guarantees of Theorem 3. We also give an application to private vector aggregation as a simple corollary of Theorem 3.

We first relate DP protocols for summation under a  $\gamma$ -imperfect shuffler to  $\sigma$ -secure protocols. Lemma 4.1 in [4] showed this relationship for uniform shufflers. It turns out their proof extends to  $\gamma$ -imperfect shufflers as well.

► **Lemma 14** (Lemma 4.1 in [4]). *Given a  $\sigma$ -secure protocol in the  $\gamma$ -I-shuffle model for  $n$ -party private summation on  $\mathbb{Z}_q$  such that each player sends  $f(n, q, \sigma)$  bits of messages, there exists an  $(\varepsilon, (1 + e^\varepsilon)2^{-\sigma-1})$ -DP protocol for any  $\varepsilon \leq O(1)$  in the  $\gamma$ -I-shuffle model for  $n$ -party private summation on real numbers with expected absolute error  $\mathcal{O}(\frac{1}{\varepsilon})$  such that each player sends  $f(n, O(n^{3/2}), \sigma)$  bits of messages.*

In Section 4, we prove the following guarantees about the split-and-mix protocol from [30].

► **Theorem 15.** *Let  $n \geq 19$  and  $\gamma \leq \frac{\log \log n}{80}$  be a distortion parameter. For worst-case statistical security with parameter  $\sigma$ , it suffices to use  $m = \mathcal{O}\left(e^{4\gamma} + \frac{e^{4\gamma}(\sigma + \log n)}{\log n}\right)$  messages, where each message has  $\mathcal{O}(\log q)$  bits, for  $q = \lceil 2n^{3/2} \rceil$ .*

By Lemma 14 and Theorem 15, we have our main statement:

► **Theorem 3.** *Let  $n \geq 19$  and  $\gamma \leq \frac{\log \log n}{80}$  be a distortion parameter. Then there exists an  $(\varepsilon, \delta)$ -DP protocol for summation in the  $\gamma$ -I-shuffle model with expected absolute error  $\mathcal{O}(\frac{1}{\varepsilon})$  and  $m = \mathcal{O}\left(e^{4\gamma} + \frac{e^{4\gamma}(\log \frac{1}{\delta} + \log n)}{\log n}\right)$  messages per party. Each message uses  $\mathcal{O}(\log q)$  bits, for  $q = \lceil 2n^{3/2} \rceil$ .*

**Applications to private vector summation.** An immediate application of our results is to the problem of private vector aggregation, where  $n$  parties have vectors  $\vec{x}_1, \dots, \vec{x}_n \in [0, 1]^d$  and the goal is to privately compute  $\vec{X} = \sum_{i=1}^n \vec{x}_i \in \mathbb{R}^d$ . Given a protocol  $\mathcal{P}$  for private summation where  $n$  players each send  $m$  messages, the  $n$  players can perform a protocol  $\mathcal{P}'$  for vector aggregation by performing  $\mathcal{P}$  on each of their  $d$  coordinates. In particular, the  $n$  players can first perform  $\mathcal{P}$  on the first coordinate of their vectors, then perform  $\mathcal{P}$  on the second coordinate of their vectors, and so on and so forth, by sending  $md$  messages in total. Equivalently, the  $n$  players can perform  $\mathcal{P}$  on a field of size  $q^d$  rather than size  $q$  and just send  $m$  messages in total. However, the total communication size is still the same, because each message increases by a factor of  $d$  due to the larger field size. Thus we consider the approach where the  $n$  players perform  $\mathcal{P}$  on each of the  $d$  coordinates.

To argue privacy, we observe that the  $n$  players run  $d$  iterations of the protocol  $\mathcal{P}$ , once for each of the coordinates. By composition of DP, i.e., Theorem 6, to guarantee  $\varepsilon$ -privacy for the overall protocol, it suffices to run each of the  $d$  iterations with privacy  $\varepsilon' = \frac{\varepsilon}{d}$  and failure probability  $\delta' = \frac{\delta}{d}$ . By post-processing of DP, i.e., Theorem 7, the resulting vector where each coordinate is computed using the corresponding protocol is  $(\varepsilon, \delta)$ -DP.

Then as a corollary to Theorem 3 with privacy parameter  $\varepsilon' = \frac{\varepsilon}{d}$  and failure probability  $\delta' = \frac{\delta}{d}$ , we obtain:



► **Theorem 16.** *Let  $n \geq 19$ ,  $d \geq 1$ ,  $\varepsilon > 0$  be a (constant) privacy parameter, and  $\gamma \leq \frac{\log \log n}{80}$  be a distortion parameter. Then there exists an  $(\varepsilon, \delta)$ -DP protocol for vector summation in the  $\gamma$ -I-shuffle model with expected absolute error  $\mathcal{O}\left(\frac{d}{\varepsilon}\right)$  per coordinate and  $m = \mathcal{O}\left(d \left(e^{4\gamma} + \frac{e^{4\gamma}(\log \frac{d}{\varepsilon} + \log n)}{\log n}\right)\right)$  messages per party. Each message uses  $\mathcal{O}(\log q)$  bits, for  $q = \lceil 2n^{3/2} \rceil$ .*

## 4 Security of Split-and-Mix Protocol

In this section, we prove the  $\sigma$ -security of the split-and-mix protocol. The proof largely attempts to follow the outline of the split-and-mix protocol analysis for private aggregation by [4], which first reduces from worst-case input to average-case input and then analyzes the connectivity of the resulting communication graph induced by a uniform shuffle.

We similarly first reduce from worst-case input to average-case input and then analyze the connectivity of the resulting communication graph induced by a uniform shuffle. The former appears in Section 4.1 and the latter appears in Section 4.2.

However, the main challenge is that the symmetric properties of the uniform shuffler is often crucially utilized in various steps of the approach. Unfortunately, these properties do not often seem to translate to  $\gamma$ -imperfect shufflers, where we might not even know the mass that is placed on each permutation. Thus we need to handle a number of technical challenges to recover qualitatively similar structural properties to the uniform shuffling model. Along the way, we show that the composition of two shufflers, where the inner shuffler is a  $\gamma$ -imperfect shuffler, is also a  $\gamma$ -imperfect shuffler with the same parameter, which can be interpreted as a post-processing statement for  $\gamma$ -imperfect shuffling.

We first formally define the split-and-mix protocol:

► **Definition 17** (Split-and-Mix Protocol, e.g., [30]). *Given an integer parameter  $m \geq 1$ , the  $m$ -message  $n$ -player split-and-mix protocol  $\mathcal{P}_{m,n}$  is defined as follows. Each player  $i$  outputs a set of  $m$  messages  $x_{i,1}, \dots, x_{i,m}$  uniformly at random conditioned on  $x_{i,1} + \dots + x_{i,m} = x_i$ . For each  $j \in [m]$ , the set of messages  $x_{1,j}, \dots, x_{n,j}$  are then swapped according to a  $\gamma$ -imperfect shuffler  $\mathcal{S}^{(j)}$ .*

### 4.1 Worst-case to Average-case Reduction

In this section, we show a reduction from worst-case input to average-case input. In other words, rather than analyze the split-and-mix protocol over the worst-case input, we show it suffices to analyze the expected performance of the split-and-mix protocol across all possible inputs. The approach is nearly identical to that of [6], but they can further simplify their final expression due to the symmetric properties of the uniform shuffler, which do not hold for the  $\gamma$ -imperfect shuffler.

Let  $\mathcal{P}_{m,n}$  denote the  $m$ -message  $n$ -player split-and-mix protocol and let  $\tilde{\mathcal{P}}_{m,n}$  be defined as follows. Each player  $i$  outputs a set of  $m+1$  messages  $x_{i,1}, \dots, x_{i,m+1}$  uniformly at random conditioned on  $x_{i,1} + \dots + x_{i,m+1} = x_i$ . For each  $i \in [n]$ , we use the notation  $\mathcal{R}_m(x_i) = (x_{i,1}, \dots, x_{i,m})$  to denote the choice of the  $m$  messages for player  $i$ . Let  $\mathbb{G} = \mathbb{F}_q$  and for  $j \in [m]$ , let  $\mathcal{S}^{(j)} : \mathbb{G}^n \rightarrow \mathbb{G}^n$  be independent shufflers. Then the output of  $\tilde{\mathcal{P}}_{m,n}$  is  $\mathcal{S}^{(j)}$  applied to the first  $m$  messages of each player, concatenated with the unshuffled final message of each player, i.e.,  $\tilde{\mathcal{P}}_{m,n}(x_1, \dots, x_n) = \mathcal{S}^{(1)}(x_{1,1}, \dots, x_{n,1}) \circ \dots \circ \mathcal{S}^{(m)}(x_{1,m}, \dots, x_{n,m}) \circ x_{1,m+1}, \dots, x_{n,m+1}$ .

We first reduce the problem to average-case statistical security using the approach of Lemma 6.1 in [6]. Formally, we say that a protocol  $\mathcal{P}_{m,n}$  provides average-case statistical security with parameter  $\sigma$  if  $\mathbb{E}_{\vec{X}, \vec{X}'}[\text{TVD}_{|\vec{X}, \vec{X}'|}(\mathcal{P}_{m,n}(\vec{X}), \mathcal{P}_{m,n}(\vec{X}'))] \leq 2^{-\sigma}$ , where  $\vec{X}$  and  $\vec{X}'$  are

## 17:10 Differentially Private Aggregation via Imperfect Shuffling

each drawn uniformly at random from all pairs of vectors in  $\mathbb{G}^n$  with the same sum. Here we use the notation  $\text{TVD}_{|\vec{x}, \vec{x}'}$  to denote the total variation distance between two distributions conditioned on fixings of  $\vec{X}$  and  $\vec{X}'$ .

► **Lemma 18.** *Suppose  $\mathcal{P}_{m,n}$  provides average-case statistical security with parameter  $\sigma$ , then  $\mathcal{P}_{m+1,n}$  and  $\tilde{\mathcal{P}}_{m,n}$  provide worst-case statistical security with parameter  $\sigma$ .*

**Proof.** Let  $\vec{x}$  and  $\vec{x}'$  be a pair of vectors in  $\mathbb{G}^n$  with the same sum. Given an output of  $\tilde{\mathcal{P}}_{m,n}(\vec{x})$ , the protocol  $\mathcal{P}_{m+1,n}(\vec{x})$  can be simulated by using an additional application of  $\mathcal{R}_{m+1}$  to randomly permute the last message of each of the players according to the distribution of the  $\gamma$ -imperfect shuffle. Hence,  $\text{TVD}(\mathcal{P}_{m+1,n}(\vec{x}), \mathcal{P}_{m+1,n}(\vec{x}')) \leq \text{TVD}(\tilde{\mathcal{P}}_{m,n}(\vec{x}), \tilde{\mathcal{P}}_{m,n}(\vec{x}'))$ . It thus suffices to upper bound the worst-case statistical security of  $\tilde{\mathcal{P}}_{m,n}$  by  $\sigma$ .

The worst-case security of  $\tilde{\mathcal{P}}_{m,n}$  is reduced to the average-case security of  $\tilde{\mathcal{P}}_{m,n}$  by noting that the addition of the  $(m+1)$ st message to each player can effectively be viewed as adding a random value to each player's input and thus transforming each input value  $x_i$  into a uniformly random value in  $\mathbb{G}$ . More formally, consider the definition  $\mathcal{R}_{m+1}(x) = (\mathcal{R}_m(x - \mathbf{U}), \mathbf{U})$ , for  $x \in \mathbb{G}$ , where  $\mathbf{U}$  is a uniformly random element of  $\mathbb{G}$ .

Since  $\vec{x} - \vec{\mathbf{U}}$  is a uniformly random vector in  $\mathbb{G}^n$ , then we can couple the randomness observed from two instances  $\vec{\mathbf{U}}, \vec{\mathbf{U}}'$  resulting from two independent executions of  $\mathcal{P}_{m,n}$  with two inputs having the same sum. Therefore,

$$\begin{aligned} \text{TVD}(\tilde{\mathcal{P}}_{m+1,n}(\vec{x}), \tilde{\mathcal{P}}_{m+1,n}(\vec{x}')) &= \text{TVD}((\mathcal{P}_{m,n}(\vec{x} - \vec{\mathbf{U}}), \vec{\mathbf{U}}), (\mathcal{P}_{m,n}(\vec{x}' - \vec{\mathbf{U}}'), \vec{\mathbf{U}}')) \\ &= \mathbb{E}_{\vec{\mathbf{U}}, \vec{\mathbf{U}}'} [\text{TVD}(\mathcal{P}_{m,n}(\vec{x} - \vec{\mathbf{U}}), \mathcal{P}_{m,n}(\vec{x}' - \vec{\mathbf{U}}'))] \\ &= \mathbb{E}_{\vec{X}, \vec{X}'} [\text{TVD}(\mathcal{P}_{m,n}(\vec{X}), \mathcal{P}_{m,n}(\vec{X}'))], \end{aligned}$$

where  $\vec{X}, \vec{X}'$  are chosen uniformly at random conditioned on  $\vec{X} = \vec{X}' + \vec{x} - \vec{x}'$ . ◀

We now upper bound the expected total variation distance between the two independent executions of the  $\gamma$ -imperfect shuffle, using an approach similar to Lemma C.1 in [6].

► **Lemma 19.** *Let  $\vec{X}$  and  $\vec{X}'$  be drawn uniformly at random from all pairs of vectors in  $\mathbb{G}^n$  with the same sum, noting that  $\vec{X}$  and  $\vec{X}'$  are not independent. For two independent executions  $\mathcal{P}_{m,n}$  and  $\mathcal{P}'_{m,n}$  of the  $\gamma$ -imperfect shuffle,  $\mathbb{E}_{\vec{X}, \vec{X}'} [\text{TVD}_{|\vec{X}, \vec{X}'}(\mathcal{P}_{m,n}(\vec{X}), \mathcal{P}_{m,n}(\vec{X}'))] \leq \sqrt{q^{mn-1} \Pr[\mathcal{P}_{m,n}(\vec{X}) = \mathcal{P}'_{m,n}(\vec{X})]} - 1$ .*

We note that the probability that two independent executions of the protocol can be decomposed into the split protocol and the mix protocol as follows. By comparison, Lemma C.2 in [6] was able to prove a simpler relationship by leveraging properties of their symmetric shuffler, which we do not have for an imperfect shuffler.

► **Lemma 20.** *Let  $\mathcal{R}_{m,n}$  and  $\mathcal{R}'_{m,n}$  denote two independent executions of the split protocol in  $\mathcal{P}_{m,n}$  so that  $\mathcal{P}_{m,n} = \mathcal{S}_{m,n} \circ \mathcal{R}_{m,n}$ . Then*

$$\Pr[\mathcal{P}_{m,n}(\vec{X}) = \mathcal{P}'_{m,n}(\vec{X})] = \Pr[\mathcal{R}_{m,n}(\vec{X}) = \mathcal{S}_{m,n}^{-1} \circ \mathcal{S}'_{m,n} \circ \mathcal{R}'_{m,n}(\vec{X})].$$

**Proof.** Note that

$$\begin{aligned} \Pr[\mathcal{P}_{m,n}(\vec{X}) = \mathcal{P}'_{m,n}(\vec{X})] &= \Pr[\mathcal{S}_{m,n} \circ \mathcal{R}_{m,n}(\vec{X}) = \mathcal{S}'_{m,n} \circ \mathcal{R}'_{m,n}(\vec{X})] \\ &= \Pr[\mathcal{R}_{m,n}(\vec{X}) = \mathcal{S}_{m,n}^{-1} \circ \mathcal{S}'_{m,n} \circ \mathcal{R}'_{m,n}(\vec{X})]. \end{aligned}$$

◀

From Lemma 19 and Lemma 20, we have

► **Lemma 21.** *For two independent executions  $\mathcal{P}_{m,n}$  and  $\mathcal{P}'_{m,n}$  of the split-and-mix protocol with a  $\gamma$ -imperfect shuffler,*

$$\mathbb{E}_{\vec{X}, \vec{X}'}[\text{TVD}(\mathcal{P}_{m,n}(\vec{X}), \mathcal{P}_{m,n}(\vec{X}'))] \leq \sqrt{q^{mn-1} \Pr \left[ \mathcal{R}_{m,n}(\vec{X}) = \mathcal{S}_{m,n}^{-1} \circ \mathcal{S}'_{m,n} \circ \mathcal{R}'_{m,n}(\vec{X}') \right]} - 1.$$

## 4.2 Reduction to Connected Components

In this section, we prove the following general statement upper bounding the probability that the shuffler  $\mathcal{S}_{m,n}^{-1} \circ \mathcal{S}'_{m,n}(\cdot)$  on the output of a randomizer achieves the same output as an independent instance of the randomizer by the expectation of a quantity relating to the number of connected components in the communication graph of the shuffler  $\mathcal{S}_{m,n}^{-1} \circ \mathcal{S}'_{m,n}(\cdot)$ . Specifically, we can view a protocol  $\mathcal{P}_{m,n}$  that is an ordered tuple  $\pi_1, \dots, \pi_m$ , where  $\pi_j$  is a permutation on  $[n]$  for each  $j \in [m]$ , so that in each round  $j \in [m]$ , user  $i \in [n]$  sends a message to user  $\pi_j(i)$ .

Then we can define the *communication graph* for the multi-message shuffle protocol  $\mathcal{P}_{m,n}$  as follows. The graph  $G$  consists of  $n$  vertices, which we associate with  $[n]$ , corresponding to the players  $[n]$  participating in the protocol  $\mathcal{P}_{m,n}$ . We add an edge between vertices  $i$  and  $j$  if player  $i$  passes one of their  $m$  messages to player  $j$ .

The following proof is the same as Lemma C.4 in [6].

► **Lemma 22.** *Let  $G$  be the graph on  $n$  vertices formed the communication graph of the shuffle  $\mathcal{S}^{-1} \circ \mathcal{S}'$ . Let  $C(G)$  be the number of connected components of  $G$ . Then  $\Pr \left[ \vec{\mathcal{R}}(\vec{X}) = \mathcal{S}^{-1} \circ \mathcal{S}' \circ \vec{\mathcal{R}}'(\vec{X}') \right] \leq \mathbb{E} \left[ q^{C(G)-mn} \right]$ .*

**Proof.** By the law of total expectation,

$$\Pr \left[ \vec{\mathcal{R}}(\vec{X}) = \mathcal{S}^{-1} \circ \mathcal{S}' \circ \vec{\mathcal{R}}'(\vec{X}') \right] = \mathbb{E} \left[ \Pr \left[ \vec{\mathcal{R}}(\vec{X}) = \mathcal{S}^{-1} \circ \mathcal{S}' \circ \vec{\mathcal{R}}'(\vec{X}') \mid \mathcal{S}, \mathcal{S}' \right] \right].$$

Thus for the graph  $G$  conditioned on  $\mathcal{S}$  and  $\mathcal{S}'$ , it suffices to show that

$$\Pr \left[ \vec{\mathcal{R}}(\vec{X}) = \mathcal{S}^{-1} \circ \mathcal{S}' \circ \vec{\mathcal{R}}'(\vec{X}') \mid \mathcal{S}, \mathcal{S}' \right] = q^{C(G)-mn}.$$

Note that  $C(G)$  depends on the choices of  $\mathcal{S}$  and  $\mathcal{S}'$  but we omit these dependencies in the notation for the sake of presentation. Recall that  $\mathcal{P}_{m,n}(\vec{X}) = \mathcal{S}_{m,n} \circ \vec{\mathcal{R}}_{m,n}(\vec{X})$  is currently indexed so that the first message of each player after the shuffle protocol completes are the first  $n$  indices, followed by the second message of each of the  $n$  players and so forth. We thus define a re-indexing permutation  $\psi : [mn] \rightarrow [mn]$  so that the  $m$  messages of the first player will be the first  $m$  indices, followed by the  $m$  messages of the second player and so forth. That is,  $\psi(j) = \lfloor \frac{j-1}{m} \rfloor + n(j-1 \bmod n) + 1$ . Let  $W, W' \in \mathbb{G}^{mn}$  be defined so that  $W_j = \psi(\vec{\mathcal{R}}(\vec{X}))_j$  and  $W'_j = \psi(\mathcal{S}^{-1} \circ \mathcal{S}' \circ \vec{\mathcal{R}}'(\vec{X}'))_j$ . The task then becomes to show that  $\Pr [W = W' \mid \mathcal{S}, \mathcal{S}'] = q^{C(G)-mn}$ . Toward that end, for each  $j \in [mn]$ , we define  $\mathcal{E}_j$  to be the event that  $W_j = W'_j$  and  $p_j = \Pr [\mathcal{E}_j \mid \mathcal{E}_1, \dots, \mathcal{E}_{j-1}]$ , so that  $\Pr [W = W' \mid \mathcal{S}, \mathcal{S}'] = \prod_{j=1}^{mn} p_j$ .

Firstly, consider the messages that are not the last message by a particular player, i.e., consider the values of  $j \in [mn]$  that are not divisible by  $m$ . Observe that conditioning on fixed values of  $\vec{X}$  and  $\vec{\mathcal{R}}'$ , as well as the events  $\mathcal{E}_1, \dots, \mathcal{E}_{j-1}$ , the value of  $W_j$  remains uniformly distributed and has probability  $q^{-1}$  of being equal to  $W'_j$ . Hence, we have  $p_j = q^{-1}$ .

For the cases where  $j$  is divisible by  $m$ , we further consider two subcases. In particular, we consider the case where  $j$  is the largest index in  $C_j$  and the case where  $j$  is not the largest index in  $C_j$ , where  $C_j$  is the set of vertices in the same connected component as  $j$  in  $G$ .

## 17:12 Differentially Private Aggregation via Imperfect Shuffling

In the first subcase, the multisets of  $W'$  and  $\vec{\mathcal{R}}'(X')$  restricted to  $C_i$  are the same and thus the multisets of the summands are the same, so that  $\sum_{i|C_i=C_j} W'_i = \sum_{i|C_i=C_j} \psi(\vec{\mathcal{R}}'(X'))_i$ . Moreover, since the indices corresponding to all messages of a fixed player are in the same connected component, then  $\sum_{i|C_i=C_j} \psi(\vec{\mathcal{R}}'(X'))_i = \sum_{i|C_i=C_j} W_i$ . Finally, we have that conditioning on  $\mathcal{E}_1, \dots, \mathcal{E}_{j-1}$  and the fact that  $j$  is the largest index in  $C_j$ ,  $\sum_{i|C_i=C_j, i \neq j} W'_i = \sum_{i|C_i=C_j, i \neq j} W_i$ . Therefore, we have  $p_j = 1$ .

For the second subcase, we shall show that  $p_j = q^{-1}$ . Let  $\mathcal{T}$  be the subset of  $(W, W') \in \mathbb{G}^{2mn}$  that are consistent with  $\mathcal{E}_1, \dots, \mathcal{E}_{j-1}$  and a fixed value of  $\vec{X}$ . We show there exists a homomorphism  $\phi : \mathbb{G} \rightarrow \mathbb{G}^{2mn}$  that maps from  $g \in \mathbb{G}$  to a  $u_g \in \mathbb{G}^{2mn}$  with a specific property to be defined. We then consider the action of  $\mathbb{G}^{2mn}$  on itself by addition of  $u_g$ . Then the property of  $\phi$  that we show is that  $u_g$  fixes  $\mathcal{T}$  and  $W_j$  but adds  $g$  to  $W'_j$ . Consider the partitioning of  $\mathcal{T}$  into equivalence classes where two elements of  $\mathcal{T}$  are equivalent if they are equal under addition by  $u_g$  for some  $g$ . Then the homomorphism induces a partitioning of  $\mathcal{T}$  into subsets of size  $q$  such that each subset contains exactly one element for which  $\mathcal{E}_j$  holds. Since each value of  $\mathcal{T}$  is equally probable, it then follows that  $p_j = q^{-1}$  as desired.

We now define the homomorphism  $\phi$  as follows. Since there exists a path in  $G$  from the vertex with the  $j$ th message to a higher index vertex, then there exists some path parameter  $\ell$  and a corresponding path  $(a_1, b_1, \dots, a_\ell, b_\ell, a_{\ell+1})$  such that the following hold. Firstly, each of the terms  $a_i, b_i$  are elements of  $[mn]$  that will ultimately map to indices of elements in  $\mathbb{G}^{mn}$ . Secondly, for all  $i \in [\ell]$ , we have  $\pi(b_i) = a_i$  for the permutation  $\pi$  induced by the  $m$  message  $n$  player protocol and moreover,  $b_i$  and  $a_{i+1}$  correspond to the same vertex. Finally, it holds that  $a_1 = j$ ,  $b_\ell > j$ ,  $a_i \neq a_{i'}$  for any  $i \neq i'$ , and  $b_i < j$  for all  $i < \ell$ . Then we implicitly define the homomorphism  $\phi$  by defining  $u_g$  to be the element of  $\mathbb{G}^{2mn}$  with the value  $g$  in the entries  $a_2, \dots, a_{\ell+1}, b_1 + mn, \dots, b_\ell + mn$  and the identity 0 in all other coordinates, where we recall that the elements  $a_i$  and  $b_i$  correspond to indices of elements in  $\mathbb{G}^{mn}$ .

We observe that the group action of addition by  $u_g$  does not affect the realization of  $\mathcal{E}_1, \dots, \mathcal{E}_{j-1}$  since  $W_{a_i}$  and  $W'_{a_i} = \vec{\mathcal{R}}'(\vec{X})_{b_i}$  are increased by exactly the same amount by  $u_g$ , except for the case when  $i = 1$  or  $i = \ell + 1$ . However, note that  $a_i \geq j$  for both of the cases where  $i = 1$  and  $i = \ell + 1$ , which does not affect the realization of  $\mathcal{E}_1, \dots, \mathcal{E}_{j-1}$ . Hence,  $u_g$  has the desired properties and so it follows that  $p_j = q^{-1}$ .

Therefore, conditioned on any fixed realization of  $\mathcal{S}$ , we have that  $\prod_{j=1}^{mn} p_j = q^{C(G)-mn}$ , so that in summary  $\Pr \left[ \vec{\mathcal{R}} = \mathcal{S}^{-1} \circ \mathcal{S}' \circ \vec{\mathcal{R}}'(\vec{X}) \right] \leq \mathbb{E}[q^{C(G)-mn}]$ .  $\blacktriangleleft$

We remark that the statement of Lemma 22 holds even for a general shuffler  $\mathcal{S}$  with the corresponding communication graph, rather than the specific shuffler  $\mathcal{S}_{m,n}^{-1} \circ \mathcal{S}'_{m,n}(\cdot)$ .

We now show that the composition of two shufflers, where the inner shuffler is a  $\gamma$ -imperfect shuffler, is also a  $\gamma$ -imperfect shuffler with the same parameter.

**► Lemma 23.** *Let  $\mathcal{S}, \mathcal{S}'$  be two shufflers such that  $\mathcal{S}$  is a  $\gamma$ -imperfect shuffler. Then,  $\mathcal{S}' \circ \mathcal{S}$  is a  $\gamma$ -imperfect shuffler.*

**Proof.** Let  $\mathcal{S}'$  be an arbitrary shuffler and  $\mathcal{S}$  be a  $\gamma$ -imperfect shuffler. Then, for any  $\pi, \pi' \in \Pi$ ,

$$\begin{aligned} \Pr[\mathcal{S}' \circ \mathcal{S} = \pi] &= \Pr[\mathcal{S} = (\mathcal{S}')^{-1} \circ \pi] \leq e^{\gamma \cdot \text{Swap}((\mathcal{S}')^{-1} \circ \pi, (\mathcal{S}')^{-1} \circ \pi')} \Pr[\mathcal{S} = (\mathcal{S}')^{-1} \circ \pi'] \\ &= e^{\gamma \cdot \text{Swap}(\pi, \pi')} \Pr[\mathcal{S}' \circ \mathcal{S} = \pi']. \end{aligned} \quad \blacktriangleleft$$

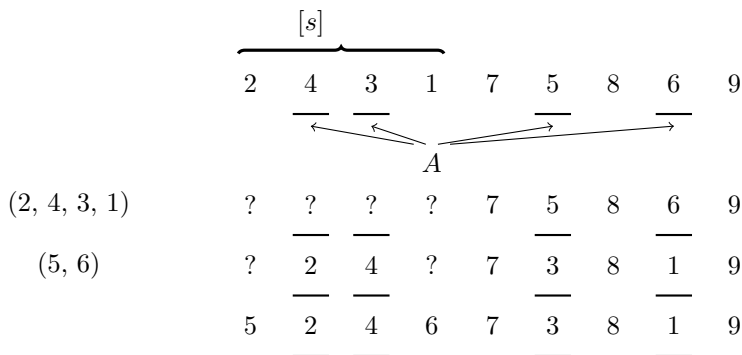
We now show a few structural statements that upper bound the probability that there exists no edge from a set  $S \subset [n]$  to  $[n] \setminus S$  for a communication graph induced by a  $\gamma$ -imperfect shuffler.

► **Lemma 24.** *Let  $G$  be the communication graph of a  $\gamma$ -imperfect shuffler (on an  $n$ -player  $m$ -message protocol). For a fixed set  $S$  of size  $s$ , the probability that there exists no edge from  $S$  to  $[n] \setminus S$  in  $G$  is at most  $e^{2sm\gamma} \binom{n}{s}^{-m}$  for  $s \leq \frac{n}{2}$  and at most  $e^{2(n-s)m\gamma} \binom{n}{s}^{-m}$  for  $s \geq \frac{n}{2}$ .*

**Proof.** Without loss of generality, let  $S = [s]$ , i.e.,  $S$  is the first  $s$  integers of  $[n]$ . Then for a permutation to not induce an edge between  $S$  and  $[n] \setminus S$ , the permutation can be decomposed into a permutation of the first  $s$  integers and a permutation of the remaining  $n - s$  integers. Hence, there are  $s!(n - s)!$  permutations of  $[n]$  such that  $S$  is preserved. Let  $\Pi_S$  be the set of permutations that preserves  $S$  so that  $|\Pi_S| = s!(n - s)!$ .

For each permutation  $\pi \in \Pi_S$ , we define a subset  $C_\pi$  of permutations so that (1)  $C_{\pi'} \cap C_\pi = \emptyset$  for all  $\pi, \pi' \in \Pi_S$  with  $\pi \neq \pi'$ , (2)  $\pi$  is the only permutation of  $C_\pi$  that preserves  $S$ , (3)  $|C_\pi| = \binom{n}{s}$ , and (4)  $\pi$  and  $\pi'$  have swap distance at most  $2s$  for any  $\pi' \in C_\pi$ , hence implying that  $\Pr[S = \pi] \leq e^{2s\gamma} \cdot \Pr[S = \pi']$ . Recall that since  $\pi \in \Pi_S$ , then  $\pi$  can be decomposed into permutations  $\pi_1$  of the first  $s$  integers and permutations  $\pi_2$  of the remaining  $n - s$  integers.

Let  $A$  be any set of  $s$  indices of  $[n]$ , sorted in increasing order. Consider the following transformation  $T_A$  on a permutation  $\pi$  to produce a permutation  $\psi$ . Place the elements of  $\pi$  in positions  $[s]$  in order into the  $s$  indices of  $A$ , so that  $\pi'(A_i) = \pi(i)$ . For the supplanted indices that have not been assigned to indices in  $A$ , place them in order into the remaining positions of  $[s]$ . Formally, let  $X = [s] \setminus A$  and  $Y = A \setminus [s]$ . Then we set  $\pi'(X_i) = \pi(Y_i)$  for all  $i \in [|X|]$ , noting that  $|X| = |Y|$ . We then define  $C_\pi$  to be the set of permutations that can be obtained from this procedure, i.e.,  $C_\pi = \{\pi' : \exists A \text{ with } \pi = T_A(\pi')\}$ . See Figure 1 for an example of the application of such an example  $T_A$ .



■ **Figure 1** An example of the transformation  $T_A$  for the permutation  $\pi = (8, 4, 6, 2, 1, 3, 7, 5, 9)$ , with  $n = 9$ ,  $s = 4$ , and  $A = (2, 3, 6, 8)$ . Note that the order  $(8, 4, 6, 2)$  is preserved within the indices of  $A$  in the resulting permutation  $\pi' = T_A(\pi)$  and the order  $(3, 5)$  is preserved within the indices  $[s] \setminus A$ .

We first claim that  $C_{\pi'} \cap C_\pi = \emptyset$  for all  $\pi, \pi' \in \Pi_S$  with  $\pi \neq \pi'$ . Suppose by way of contradiction, there exists  $\psi \in C_\pi \cap C_{\pi'}$ , so that there exist sets  $A$  and  $A'$  with  $\psi = T_A(\pi) = T_{A'}(\pi')$ . Recall that since  $\pi, \pi' \in \Pi_S$ , then  $\pi, \pi'$  can be decomposed into permutations  $\pi_1, \pi'_1$  of the first  $s$  integers and permutations  $\pi_2, \pi'_2$  of the remaining  $n - s$  integers. After applying  $T_A$  to  $\pi$ , then the first  $s$  integers are in the indices of  $A'$ , in some order. Similarly, after applying  $T_{A'}$  to  $\pi'$ , then the first  $s$  integers are in the indices of  $A$ , in some order. Hence for  $\psi = T_A(\pi) = T_{A'}(\pi')$ , it follows that  $A = A'$ , so it suffices to show that  $T_A$  is injective for a fixed  $A$ . To that end, note that  $T_A$  preserves the order of  $[s]$  within  $A$  and thus if  $\pi = \pi_1 \circ \pi_2$ , then  $\pi_1$  is the restriction of  $T_A(\pi)$  to  $A$ . Similarly, note that  $T_A$  does not touch the indices outside of  $A \cup [s]$  and so  $\pi_2$  is preserved by  $T_A(\pi)$  in the restriction of  $[n] \setminus (A \cup [s])$ . Finally,

## 17:14 Differentially Private Aggregation via Imperfect Shuffling

$T_A$  preserves the relative order of  $\pi_2$  inside the indices of  $[s] \setminus A$ . Therefore, given  $A$  and  $T_A(\pi)$ , we can completely recover  $\pi_1$  and  $\pi_2$  and thus  $\pi$ . In other words,  $T_A$  is injective, so that  $T_A(\pi) = T_A(\pi')$  implies  $\pi = \pi'$ , which is a contradiction.

To see that  $\pi$  is the only permutation of  $C_\pi$  that preserves  $S$ , note that if any of the  $s$  positions are picked outside  $[s]$ , then there exists a value of  $[s]$  outside of the first  $s$  positions and so the resulting permutation does not preserve  $S$ . However, there is only a single way to pick  $s$  indices from  $[n]$  that are all inside  $[s]$ , which corresponds to  $\pi$ . Hence,  $\pi$  is the only permutation of  $C_\pi$  that preserves  $S$ .

To see the third property, note that  $A$  is formed by choosing  $s$  indices of  $[n]$ . Hence,  $|A| = \binom{n}{s}$ . Since  $A$  is exactly the set of positions for which  $\pi_1$  is mapped to, then each element of  $A$  corresponds to a unique element in  $C_\pi$ . Thus,  $|C_\pi| = \binom{n}{s}$ .

To see the fourth property, note that the only swaps are indices in  $A$  with indices in  $[s]$ , meaning that at most  $2s$  indices are changed. Thus,  $\pi$  and  $\pi'$  have swap distance at most  $2s$  for any  $\pi' \in C_\pi$ . Then by the  $\gamma$ -imperfect shuffle property,  $\Pr[\mathcal{S} = \pi] \leq e^{2s\gamma} \cdot \Pr[\mathcal{S} = \pi']$ .

Since we have associated each  $\pi \in \Pi_S$  with a set  $C_\pi$  of size  $\binom{n}{s}$  such that  $\pi' \notin C_\pi$  for  $\pi' \in \Pi_S$  with  $\pi' \neq \pi$  and  $\Pr[\mathcal{S} = \pi] \leq e^{s\gamma} \cdot \Pr[\mathcal{S} = \pi']$ , then it follows from a coupling argument that the probability that there exists no edge from  $S$  to  $[n] \setminus S$  after one iteration of the  $\gamma$ -imperfect shuffle is at most  $e^{2s\gamma} \binom{n}{s}^{-1}$ . By independence, the probability that there exists no edge from  $S$  to  $[n] \setminus S$  in  $G$  after  $m$  iterations is at most  $e^{2sm\gamma} \binom{n}{s}^{-m}$ .

By symmetry for sets  $S$  with size  $s$  and  $n - s$ , we have the probability is at most  $\min\left(e^{2sm\gamma} \binom{n}{s}^{-m}, e^{2(n-s)m\gamma} \binom{n}{s}^{-m}\right)$  across all ranges of  $s$ .  $\blacktriangleleft$

► **Lemma 25.** *Let  $G$  be the communication graph of a  $\gamma$ -imperfect shuffler (on an  $n$ -player  $m$ -message protocol). For a fixed set  $S$  with size  $s$ , the probability that there exists no edge from  $S$  to  $[n] \setminus S$  in  $G$  is at most  $e^{km\gamma} \binom{n/2}{k}^{-m}$  for any integer  $k$  with  $0 \leq k \leq \min(s, n - s)$ .*

By Lemma 24 and Lemma 25, we have:

► **Lemma 26.** *Let  $G$  be the communication graph of a  $\gamma$ -imperfect shuffler (on an  $n$ -player  $m$ -message protocol). For a fixed set  $S$  with size  $s$ , the probability that there exists no edge from  $S$  to  $[n] \setminus S$  in  $G$  is at most  $e^{2sm\gamma} \binom{n}{s}^{-m}$  for  $s \leq \frac{n}{2}$ , at most  $e^{2(n-s)m\gamma} \binom{n}{s}^{-m}$  for  $s \geq \frac{n}{2}$ , and at most  $e^{km\gamma} \binom{n/2}{k}^{-m}$  for any integer  $k$  with  $0 \leq k \leq \min(s, n - s)$ .*

Lemma 23 and Lemma 26 are the two main structural properties of imperfect shufflers that we use to overcome the challenge of adapting the analysis of [6] to shufflers without symmetry.

We now upper bound the probability that the number of connected components of  $G$  is  $c$ , where  $G$  is the underlying communication graph for the split-and-mix-protocol under a  $\gamma$ -imperfect shuffle.

► **Lemma 27.** *Let  $n \geq 19$  and  $m \geq 8e^{4\gamma}$ . Let  $G$  be the communication graph of a  $\gamma$ -imperfect shuffler (on an  $n$ -player  $m$ -message protocol). Let  $p(n, c)$  denote the probability that the number of connected components of  $G$  is  $c$ . Then  $p(n, c) \leq \frac{2^{c-1}}{c!} \left(\frac{e}{n}\right)^{\frac{(m-1)(c-1)}{32e^{4\gamma}}} \cdot e^{2\gamma(m-1)(c-1)}$ .*

We now upper bound the expected value of  $\mathbb{E}[q^{C(G)}]$  for the purposes of upper bounding the right hand side of Lemma 22.

► **Lemma 28.** *Let  $n \geq 19$ ,  $m \geq 8e^{4\gamma}$ , and  $q \leq \left(\frac{n}{e}\right)^{\frac{(m-1)}{32e^{4\gamma}}} e^{2\gamma(1-m)}$ . Let  $G$  be the graph on  $n$  vertices formed a random instantiation of the split-and-mix protocol  $\mathcal{P}_{m,n}$  with  $m$  messages for each of  $n$  players, using a  $\gamma$ -imperfect shuffler  $\mathcal{S}$ . That is, let  $G$  have an edge between  $i$  and  $j$  if and only if player  $i$  passes one of their  $m$  messages to player  $j$ . Then  $\mathbb{E}[q^{C(G)}] \leq q + 3q^2 e^{2\gamma(m-1)} \left(\frac{e}{n}\right)^{\frac{m-1}{32e^{4\gamma}}}$ .*



**Proof.** By Lemma 27, we have  $p(n, c) \leq \frac{2^{c-1}}{c!} \left(\frac{\epsilon}{n}\right)^{\frac{(m-1)(c-1)}{32e^{4\gamma}}} \cdot e^{2\gamma(m-1)(c-1)}$ . Taking the expectation, we have  $\mathbb{E}[q^{C(G)}] \leq \sum_{c=1}^n q^c \frac{2^{c-1}}{c!} \left(\frac{\epsilon}{n}\right)^{\frac{(m-1)(c-1)}{32e^{4\gamma}}} \cdot e^{2\gamma(m-1)(c-1)}$ . Since term in the summand after the second term is at most  $\frac{2q}{3} \left(\frac{\epsilon}{n}\right)^{\frac{(m-1)}{32e^{4\gamma}}} e^{2\gamma(m-1)}$  times the previous term in the summand, then  $\mathbb{E}[q^{C(G)}] \leq q + q^2 e^{2\gamma(m-1)} \left(\frac{\epsilon}{n}\right)^{\frac{m-1}{32e^{4\gamma}}} \sum_{i=0}^{\infty} \left(\frac{2q}{3} \left(\frac{\epsilon}{n}\right)^{\frac{(m-1)}{32e^{4\gamma}}} e^{2\gamma(m-1)}\right)^i$ . Since  $q \leq \left(\frac{n}{e}\right)^{\frac{(m-1)}{32e^{4\gamma}}} e^{2\gamma(1-m)}$  by assumption, then  $\mathbb{E}[q^{C(G)}] \leq q + 3q^2 e^{2\gamma(m-1)} \left(\frac{\epsilon}{n}\right)^{\frac{m-1}{32e^{4\gamma}}}$ . ◀

We now analyze the statistical security of the split-and-mix protocol.

► **Lemma 29.** *Let  $n \geq 19$ ,  $m \geq 8e^{4\gamma}$ , and  $q \leq \left(\frac{n}{e}\right)^{\frac{(m-1)}{32e^{4\gamma}}} e^{2\gamma(1-m)}$ . Then we have worst-case statistical security with parameter  $\sigma \leq (m-1) \left(\frac{\log n - \log \epsilon}{64e^{4\gamma}} - 2\gamma \log e\right) - 3 \log(3q)$ ,*

**Proof.** By Lemma 21 and Lemma 22, we have

$$\mathbb{E}_{\vec{X}, \vec{X}'}[\text{TVD}(\mathcal{P}_{m,n}(\vec{X}), \mathcal{P}_{m,n}(\vec{X}'))] \leq \sqrt{q^{mn-1} \mathbb{E}[q^{C(G)-mn}] - 1},$$

where  $C(G)$  is the communication graph for the shuffle  $\mathcal{S}^{-1} \circ \mathcal{S}'$ . By Lemma 23 and the fact that  $\mathcal{S}'$  is a  $\gamma$ -imperfect shuffler, we have that  $\mathcal{S}^{-1} \circ \mathcal{S}'$  is also a  $\gamma$ -imperfect shuffler and thus it suffices to upper bound  $\mathbb{E}[q^{C(G)-mn}]$  where  $C(G)$  is the communication graph for an arbitrary  $\gamma$ -imperfect shuffler  $\mathcal{S}$ . Therefore by Lemma 28, we have average case statistical security less than or equal to  $2^{-\sigma} \geq \sqrt{3q^3 e^{2\gamma(m-1)} \left(\frac{\epsilon}{n}\right)^{\frac{m-1}{32e^{4\gamma}}}}$ , which holds for  $\sigma \leq (m-1) \left(\frac{\log n - \log \epsilon}{64e^{4\gamma}} - 2\gamma \log e\right) - 3 \log(3q)$ . The claim then follows by the reduction of worst-case input to average-case input by Lemma 18. ◀

Now it can be verified that by restricting  $\gamma \leq \frac{\log \log n}{80}$ , then we have both  $728e^{4\gamma} \leq \log n$  and  $\lceil 2n^{3/2} \rceil \leq \left(\frac{n}{e}\right)^{\frac{(m-1)}{32e^{4\gamma}}} e^{2\gamma(1-m)}$ . These conditions imply that 1)  $\left(\frac{\log n - \log \epsilon}{64e^{4\gamma}} - 2\gamma \log e\right) = \mathcal{O}\left(\frac{\log n}{e^{4\gamma}}\right)$ , so that the parameter  $\sigma$  has a non-empty range in the statement of Lemma 29, and 2)  $q = \lceil 2n^{3/2} \rceil$  satisfies  $q \leq \left(\frac{n}{e}\right)^{\frac{(m-1)}{32e^{4\gamma}}} e^{2\gamma(1-m)}$  in the statement of Lemma 29. As a corollary, we obtain the following guarantees for worst-case statistical security:

► **Theorem 15.** *Let  $n \geq 19$  and  $\gamma \leq \frac{\log \log n}{80}$  be a distortion parameter. For worst-case statistical security with parameter  $\sigma$ , it suffices to use  $m = \mathcal{O}\left(e^{4\gamma} + \frac{e^{4\gamma}(\sigma + \log n)}{\log n}\right)$  messages, where each message has  $\mathcal{O}(\log q)$  bits, for  $q = \lceil 2n^{3/2} \rceil$ .*

## 5 Conclusion and Discussion

In this work, we introduce the imperfect shuffle DP model, as a means of abstracting out real-world scenarios that prevent perfect shuffling. We also give a real summation protocol with nearly optimal error and small communication complexity. The protocol, which is based on the split-and-mix protocol [30], is similar to that of the (perfect) shuffle model [6, 25], while the main challenge comes in the analysis. Although we overcome this hurdle for this particular protocol, our techniques are quite specific. Therefore, an interesting open question is whether there is a general theorem that transfer the privacy guarantee in the perfect shuffle model to that in the imperfect shuffle model, possibly with some loss in the privacy parameters.



## References

- 1 Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *CCS*, pages 308–318, 2016.
- 2 John M Abowd. The US census bureau adopts differential privacy. In *KDD*, pages 2867–2867, 2018.
- 3 Naman Agarwal, Ananda Theertha Suresh, Felix X. Yu, Sanjiv Kumar, and Brendan McMahan. cpSGD: communication-efficient and differentially-private distributed SGD. In *NeurIPS*, pages 7575–7586, 2018.
- 4 Borja Balle, James Bell, Adrià Gascón, and Kobbi Nissim. Differentially private summation with multi-message shuffling. *CoRR*, abs/1906.09116, 2019. [arXiv:1906.09116](https://arxiv.org/abs/1906.09116).
- 5 Borja Balle, James Bell, Adrià Gascón, and Kobbi Nissim. The privacy blanket of the shuffle model. In *CRYPTO*, pages 638–667, 2019.
- 6 Borja Balle, James Bell, Adrià Gascón, and Kobbi Nissim. Private summation in the multi-message shuffle model. In *CCS*, pages 657–676, 2020.
- 7 Borja Balle, Peter Kairouz, Brendan McMahan, Om Dipakbhai Thakkar, and Abhradeep Thakurta. Privacy amplification via random check-ins. In *NeurIPS*, 2020.
- 8 Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *FOCS*, pages 464–473, 2014.
- 9 Amos Beimel, Kobbi Nissim, and Eran Omri. Distributed private data analysis: Simultaneously solving how and what. In *CRYPTO*, pages 451–468, 2008.
- 10 Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan R. Ullman. CoinPress: Practical private mean and covariance estimation. In *NeurIPS*, 2020.
- 11 Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnés, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *SOSP*, pages 441–459, 2017.
- 12 T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Optimal lower bound for differentially private multi-party aggregation. In *ESA*, pages 277–288, 2012.
- 13 Xiangyi Chen, Zhiwei Steven Wu, and Mingyi Hong. Understanding gradient clipping in private SGD: A geometric perspective. In *NeurIPS*, 2020.
- 14 Albert Cheu, Adam D. Smith, Jonathan R. Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *EUROCRYPT*, pages 375–403, 2019.
- 15 Albert Cheu and Chao Yan. Necessary conditions in multi-server differential privacy. In *ITCS*, pages 36:1–36:21, 2023.
- 16 Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *NIPS*, pages 3571–3580, 2017.
- 17 Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, pages 486–503, 2006.
- 18 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.
- 19 Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.
- 20 Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *SODA*, pages 2468–2479, 2019.
- 21 Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In *CCS*, pages 1054–1067, 2014.
- 22 Vitaly Feldman, Audra McMillan, and Kunal Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *FOCS*, pages 954–964, 2021.
- 23 Vitaly Feldman, Audra McMillan, and Kunal Talwar. Stronger privacy amplification by shuffling for renyi and approximate differential privacy. In *SODA*, pages 4966–4981, 2023.

- 24 Badih Ghazi, Ravi Kumar, Pasin Manurangsi, Rasmus Pagh, and Amer Sinha. Differentially private aggregation in the shuffle model: Almost central accuracy in almost a single message. In *ICML*, pages 3692–3701, 2021.
- 25 Badih Ghazi, Pasin Manurangsi, Rasmus Pagh, and Ameya Velingker. Private aggregation from fewer anonymous messages. In *EUROCRYPT*, pages 798–827, 2020.
- 26 Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM J. Comput.*, 41(6):1673–1693, 2012.
- 27 Antonious M. Girgis, Deepesh Data, Suhas N. Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of federated learning: Privacy, accuracy and communication trade-offs. *IEEE J. Sel. Areas Inf. Theory*, 2(1):464–478, 2021.
- 28 S. Dov Gordon, Jonathan Katz, Mingyu Liang, and Jiayu Xu. Spreading the privacy blanket: Differentially oblivious shuffling for differential privacy. In *ACNS*, pages 501–520, 2022.
- 29 Andy Greenberg. Apple’s “differential privacy” is about collecting your data – but not your data, June 2016.
- 30 Yuval Ishai, Eyal Kushilevitz, Rafail Ostrovsky, and Amit Sahai. Cryptography from anonymity. In *FOCS*, pages 239–248, 2006.
- 31 Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D’Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrede Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1-2):1–210, 2021.
- 32 Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, 2011.
- 33 Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *CoRR*, abs/1610.05492, 2016. [arXiv:1610.05492](https://arxiv.org/abs/1610.05492).
- 34 Ios Kotsogiannis, Yuchao Tao, Xi He, Maryam Fanaeepour, Ashwin Machanavajjhala, Michael Hay, and Gerome Miklau. Privatesql: A differentially private SQL query engine. *Proc. VLDB Endow.*, 12(11):1371–1384, 2019.
- 35 Stephen Shankland. How google tricks itself to protect chrome user privacy. *CNET*, October, 2014.
- 36 Elaine Shi and Ke Wu. Non-interactive anonymous router. In *EUROCRYPT*, pages 489–520, 2021.
- 37 Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. Stochastic gradient descent with differentially private updates. In *GlobalSIP*, pages 245–248, 2013.
- 38 Uri Stemmer. Locally private  $k$ -means clustering. *J. Mach. Learn. Res.*, 22:176:1–176:30, 2021.
- 39 Uri Stemmer and Haim Kaplan. Differentially private  $k$ -means with constant multiplicative error. In *NeurIPS*, pages 5436–5446, 2018.
- 40 Ananda Theertha Suresh, Felix X. Yu, Sanjiv Kumar, and H. Brendan McMahan. Distributed mean estimation with limited communication. In *ICML*, pages 3329–3337, 2017.
- 41 Martin Thomson and Christopher A. Wood. Oblivious http, 2023. URL: <https://datatracker.ietf.org/doc/draft-ietf-ohai-ohhttp/>.
- 42 Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *JASA*, 60(309):63–69, 1965.

- 43 Royce J. Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. Differentially private SQL with bounded user contribution. *PoPETs*, 2020(2):230–250, 2020.
- 44 Mingxun Zhou and Elaine Shi. The power of the differentially oblivious shuffle in distributed privacy mechanisms. *IACR Cryptol. ePrint Arch.*, page 177, 2022.
- 45 Mingxun Zhou, Elaine Shi, T.-H. Hubert Chan, and Shir Maimon. A theory of composition for differential obliviousness. In *EUROCRYPT*, 2023.

## A Missing Proofs

**Proof of Lemma 19.** We write  $\mathcal{P}$  and  $\mathcal{P}'$  as shorthand for  $\mathcal{P}_{m,n}$  and  $\mathcal{P}'_{m,n}$ , respectively. Let  $\vec{V}$  be a uniformly random vector drawn from  $\mathbb{G}^{mn}$ , conditioned on  $\vec{V}$  having the same sum as  $\vec{X}$  and  $\vec{X}'$ . Then by the triangle inequality,

$$\begin{aligned} \mathbb{E}_{\vec{X}, \vec{X}'} [\text{TVD}_{|\vec{X}, \vec{X}'}(\mathcal{P}(\vec{X}), \mathcal{P}(\vec{X}'))] &\leq \mathbb{E}_{\vec{X}, \vec{X}'} [\text{TVD}_{|\vec{X}, \vec{X}'}(\mathcal{P}(\vec{X}), \vec{V}) + \text{TVD}_{|\vec{X}, \vec{X}'}(\vec{V}, \mathcal{P}(\vec{X}'))] \\ &= \mathbb{E}_{\vec{X}} [\text{TVD}_{|\vec{X}}(\mathcal{P}(\vec{X}), \vec{V})] + \mathbb{E}_{\vec{X}'} [\text{TVD}_{|\vec{X}'}(\vec{V}, \mathcal{P}(\vec{X}'))] \\ &= 2\mathbb{E}_{\vec{X}} [\text{TVD}_{|\vec{X}}(\mathcal{P}(\vec{X}), \vec{V})]. \end{aligned}$$

Moreover, considering the distribution over  $\vec{V}$ ,

$$\begin{aligned} 2\text{TVD}_{|\vec{X}}(\mathcal{P}(\vec{X}), \vec{V}) &= \sum_{\vec{v} \in \mathbb{G}^{mn}} \left| \Pr[\mathcal{P}(\vec{X}) = \vec{v}] - \Pr[\vec{V} = \vec{v}] \right| \\ &= \sum_{\vec{v} \in \mathbb{G}^{mn}, \sum \vec{v} = \sum \vec{X}} \left| \Pr[\mathcal{P}(\vec{X}) = \vec{v}] - q^{1-mn} \right| \\ &= q^{mn-1} \mathbb{E}_{\vec{V}} \left[ \left| \Pr[\mathcal{P}(\vec{X}) = \vec{V}] - q^{1-mn} \right| \right]. \end{aligned}$$

Since  $\vec{V}$  is a uniformly random vector from  $\mathbb{G}^{mn}$  with its sum being equal to that of  $\vec{X}$ , then for the random variable  $\mathcal{Z} := \mathcal{Z}(\mathcal{X}, \mathcal{V}) := \Pr[\mathcal{P}(\vec{X}) = \vec{V}]$ , we have  $\mathbb{E}[\mathcal{Z}] = q^{1-mn}$ . Therefore,  $2\text{TVD}_{|\vec{X}}(\mathcal{P}(\vec{X}), \vec{V}) \leq q^{mn-1} \mathbb{E}[|\mathcal{Z} - \mathbb{E}[\mathcal{Z}]|]$ . By convexity,  $\mathbb{E}[|\mathcal{Z} - \mathbb{E}[\mathcal{Z}]|] \leq \sqrt{\mathbb{E}[\mathcal{Z}^2]}$ . Since we have

$$\mathbb{E}_{\vec{V}}[\mathcal{Z}^2] = q^{1-mn} \sum_{\vec{v} \in \mathbb{G}^{mn}, \sum \vec{v} = \sum \vec{X}} \Pr[\mathcal{P}(\vec{X}) = \vec{v}]^2 = q^{1-mn} \Pr[\mathcal{P}(\vec{X}) = \mathcal{P}'(\vec{X})],$$

we thus have

$$\begin{aligned} \mathbb{E}_{\vec{X}, \vec{X}'} [\text{TVD}_{|\vec{X}, \vec{X}'}(\mathcal{P}(\vec{X}), \mathcal{P}(\vec{X}'))] &\leq 2\text{TVD}_{|\vec{X}}(\mathcal{P}(\vec{X}), \vec{V}) \\ &\leq q^{mn-1} \mathbb{E}_{\mathcal{V}(\vec{X})} [|\Pr[\mathcal{P}(\vec{X}) = \vec{V}] - q^{1-mn}|] \\ &\leq \sqrt{q^{mn-1} \Pr[\mathcal{P}_{m,n}(\vec{X}) = \mathcal{P}'_{m,n}(\vec{X})]} - 1. \quad \blacktriangleleft \end{aligned}$$

**Proof of Lemma 25.** We can similarly show that the probability that there exists no edge from  $S$  to  $[n] \setminus S$  in  $G$  after the  $m$  iterations is at most  $e^{km\gamma} \binom{n/2}{k}^{-m}$  for any integer  $k$  with  $0 \leq k \leq \min(s, n-s)$  by the following modifications to the coupling argument. We again let  $S = [s]$  without loss of generality and let  $k \leq \min(s, n-s)$  be a fixed non-negative integer.

Recall that there are  $s!(n-s)!$  permutations of  $[n]$  such that  $S$  is preserved. We define  $\Pi_S$  to be the set of permutations that preserves  $S$  so that  $|\Pi_S| = s!(n-s)!$  and we define a transformation  $T_A(\pi)$  for a permutation  $\pi \in \Pi_S$  as follows.

If  $s \leq \frac{n}{2}$ , we let  $A$  be a set of  $k$  positions in  $\{s+1, \dots, n\}$ , sorted in increasing order. We then initialize  $\psi = \pi$  and iteratively perform the following procedure  $k$  times. For each  $i \in [k]$ , we swap the value in the  $i$ th index of  $\psi$  with the value in the  $A_i$ th index of  $A$ . We then output set  $T_A(\pi)$  to be the result of  $\psi$  after applying these  $k$  swaps. Note that since  $[s]$  and  $A$  are disjoint, we can also explicitly define the resulting  $\psi = T_A(\pi)$  by

$$\psi(i) = \begin{cases} \pi(i), & i \notin (A \cup [k]) \\ \pi(A_i), & i \in [k] \\ \pi(j), & j = A_i, i \in [k] \end{cases}.$$

Similarly, if  $s \geq \frac{n}{2}$ , we let  $A$  be a set of  $k$  positions in  $[n-s]$ , sorted in increasing order, and initialize  $\psi = \pi$ . Then for each  $i \in [k]$ , we swap the value in the  $(n-i+1)$ st index of  $\psi$  with the value in the  $i$ th index of  $A$ . Alternatively, we can also explicitly define the resulting  $\psi = T_A(\pi)$  by

$$\psi(i) = \begin{cases} \pi(i), & i \notin (A \cup \{n-k+1, \dots, n\}) \\ \pi(A_i), & i \in \{n-k+1, \dots, n\} \\ \pi(j), & j = A_i, i \in [k] \end{cases}.$$

We again define  $C_\pi$  to be the set of permutations that can be obtained from this procedure, i.e.,  $C_\pi = \{\pi' : \exists A \text{ with } \pi = T_A(\pi')\}$ . By the same argument as in Lemma 25, we have (1)  $C_{\pi'} \cap C_\pi = \emptyset$  for all  $\pi, \pi' \in \Pi_S$  with  $\pi \neq \pi'$ , (2)  $\pi$  is the only permutation of  $C_\pi$  that preserves  $S$ , (3)  $|C_\pi| = \binom{n}{k}$ . By the construction of  $T_A$  performing  $k$  swaps on  $\pi$ , we also have that  $\pi$  and  $\pi'$  have swap distance at most  $k$  for any  $\pi' \in C_\pi$ , so that  $\Pr[S = \pi] \leq e^{k\gamma} \cdot \Pr[S = \pi']$ .

Also by construction, we have  $|C_\pi| \geq \binom{n}{k}$  and so by adapting the above coupling argument, we have that the probability that there exists no edge from  $S$  to  $[n] \setminus S$  in  $G$  after the  $m$  iterations is at most  $e^{km\gamma} \binom{n/2}{k}^{-m}$ . ◀

**Proof of Lemma 27.** For a fixed set  $S$ , let  $\mathbb{P}_S$  denote the probability that there is no edge from  $S$  to  $[n] \setminus S$ . Let  $p(n, c)$  denote the probability that the number of connected components of  $G$  is  $c$ . Then

$$p(n, c) \leq \frac{1}{c} \sum_{S \subseteq [n]} \mathbb{P}_S \cdot p(n - |S|, c - 1) \leq \frac{1}{c} \sum_{s=1}^{n-c+1} \binom{n}{s} \mathbb{P}_S \cdot p(n - |S|, c - 1).$$

We decompose this sum and apply Lemma 26.

By Lemma 26, we have  $\mathbb{P}_S \leq \min(e^{2(n-s)m\gamma} \binom{n}{s}^{-m}, e^{2sm\gamma} \binom{n}{s}^{-m})$ . By Lemma 26, we also have  $\mathbb{P}_S \leq e^{km\gamma} \binom{n/2}{k}^{-m}$  for any  $k \leq \min(s, n-s)$ . Observe that for  $k \geq n-s \geq \frac{n}{2}$ , we have  $e^{2(n-s)m\gamma} \binom{n}{s}^{-m} \leq e^{2km\gamma} \binom{n}{k}^{-m} \leq e^{2km\gamma} \binom{n/2}{k}^{-m}$ . Thus for  $k = \frac{n}{4e^{4\gamma}}$ ,

$$\begin{aligned} p(n, c) &\leq \frac{1}{c} \sum_{s=1}^k \binom{n}{s} \binom{n}{s}^{-m} e^{2sm\gamma} \cdot p(n - |S|, c - 1) \\ &\quad + \frac{1}{c} \sum_{s=k+1}^{n-c+1} \binom{n}{s} \binom{n/2}{k}^{-m} e^{2km\gamma} \cdot p(n - |S|, c - 1). \end{aligned}$$

## 17:20 Differentially Private Aggregation via Imperfect Shuffling

Observe that  $k = \frac{n}{4e^{4\gamma}}$  implies that

$$\begin{aligned} e^{2\gamma} &\leq \left(\frac{n}{2k}\right)^{1/2} \\ e^{2km\gamma} &\leq \left(\frac{n}{2k}\right)^{km/2} \leq \left(\frac{n/2}{k}\right)^{m/2} \\ \left(\frac{n/2}{k}\right)^{-m} e^{2km\gamma} &\leq \left(\frac{n/2}{k}\right)^{-m/2} \leq \binom{n}{k}^{-m/2}. \end{aligned}$$

Thus we have

$$\begin{aligned} p(n, c) &\leq \frac{1}{c} \sum_{s=1}^k \binom{n}{s} \binom{n}{s}^{-m} e^{2sm\gamma} \cdot p(n - |S|, c - 1) \\ &\quad + \frac{1}{c} \sum_{s=k+1}^{n-c+1} \binom{n}{s} \binom{n}{k}^{-m/2} \cdot p(n - |S|, c - 1). \end{aligned}$$

Since  $k = \frac{n}{4e^{4\gamma}}$ , then

$$\binom{n}{k}^{-m/2} \leq (4e^{4\gamma})^{-\frac{nm}{8e^{4\gamma}}} \leq (2e)^{-\frac{nm}{8e^{4\gamma}}} \leq \left(\frac{n}{n/2}\right)^{-\frac{m}{4e^{4\gamma}}} \leq \binom{n}{s}^{-\frac{m}{4e^{4\gamma}}}.$$

Hence,

$$\begin{aligned} p(n, c) &\leq \frac{1}{c} \sum_{s=1}^k \binom{n}{s} \binom{n}{s}^{-m} e^{2sm\gamma} \cdot p(n - |S|, c - 1) \\ &\quad + \frac{1}{c} \sum_{s=k+1}^{n-c+1} \binom{n}{s}^{1-\frac{m}{4e^{4\gamma}}} \cdot p(n - |S|, c - 1). \end{aligned}$$

For  $m \geq 8e^{4\gamma}$ , we have  $1 \leq \frac{m}{8e^{4\gamma}}$  and thus

$$\begin{aligned} p(n, c) &\leq \frac{1}{c} \sum_{s=1}^k \binom{n}{s} \binom{n}{s}^{-m} e^{2sm\gamma} \cdot p(n - |S|, c - 1) \\ &\quad + \frac{1}{c} \sum_{s=k+1}^{n-c+1} \binom{n}{s}^{-\frac{m}{8e^{4\gamma}}} \cdot p(n - |S|, c - 1). \end{aligned}$$

We first apply the induction hypothesis that  $p(n, c) \leq \frac{2^{c-1}}{c!} \left(\frac{e}{n}\right)^{\frac{(m-1)(c-1)}{32e^{4\gamma}}} \cdot e^{\gamma(m-1)(c-1)}$ :

$$\begin{aligned} p(n, c) &\leq \frac{2^{c-1}}{c!} \left(\frac{e}{n}\right)^{\frac{(m-1)(c-1)}{32e^{4\gamma}}} \cdot e^{2\gamma(m-1)(c-1)} \cdot \frac{1}{2} \cdot e^{\frac{(1-m)}{32e^{4\gamma}}} \cdot e^{2\gamma(1-m)} \\ &\quad \cdot \left( \sum_{s=1}^k \binom{n}{s}^{1-m} e^{2sm\gamma} \left(\frac{n^{c-1}}{(n-s)^{c-2}}\right)^{\frac{m-1}{32e^{4\gamma}}} + \sum_{s=k+1}^{n-c+1} \left(\frac{(n-s)!s!n^{c-1}}{n!(n-s)^{c-2}}\right)^{\frac{m-1}{32e^{4\gamma}}} \right). \end{aligned}$$

We upper bound  $p(n, c)$  by upper bounding the summation across the first  $k$  terms, i.e., the head of the summation, then upper bounding the tail terms of the summation, i.e., the terms with  $s \geq \frac{3n}{4}$ , and finally upper bounding the remaining terms of the summation, i.e.,  $s \in [k, \frac{3n}{4}]$ .

**Upper bounding the head terms in the summation.** We now upper bound the summation across all  $s \leq k$ . Let  $a_s = \binom{n}{s}^{1-m} e^{2sm\gamma} \left( \frac{n^{c-1}}{(n-s)^{c-2}} \right)^{\frac{m-1}{32e^{4\gamma}}}$ . For  $s \leq k = \frac{n}{4e^{4\gamma}}$  and  $m \geq 8e^{4\gamma}$ ,

$$\begin{aligned} \frac{a_s}{a_{s-1}} &= \left( \frac{s}{n-s+1} \right)^{m-1} e^{2m\gamma} \left( \frac{n-s+1}{n-s} \right)^{\frac{(m-1)(c-2)}{32e^{4\gamma}}} \\ &\leq \left( \frac{1}{8e^{4\gamma}} \right)^{m-1} e^{2m\gamma} e^{\frac{(m-1)(c-2)}{n-s}} \\ &\leq \left( \frac{1}{8e^{4\gamma}} \right)^{m-1} (e^{4\gamma})^{m-1} e^{\frac{4(m-1)}{3}} \\ &\leq \left( \frac{e^{4/3}}{8} \right)^{m-1} \leq \left( \frac{1}{2} \right)^{m-1} \leq \frac{1}{25}. \end{aligned}$$

Then through a geometric series, we bound the summation

$$\begin{aligned} \sum_{s=1}^k a_s &\leq \sum_{s=1}^{\infty} \frac{a_1}{25^{s-1}} \leq \frac{26a_1}{25} \\ &\leq \frac{26}{25} n^{1-m} e^{m\gamma} \left( \frac{n^{c-1}}{(n-1)^{c-2}} \right)^{\frac{m-1}{32e^{4\gamma}}} \\ &\leq \frac{26}{25} e^{m\gamma} e^{\frac{m-1}{32e^{4\gamma}}} \end{aligned}$$

**Upper bounding the tail terms in the summation.** We now upper bound the summation across all  $s \geq \lceil \frac{3n}{4} \rceil$ . Let  $b_s = \left( \frac{(n-s)!s!n^{c-1}}{n!(n-s)^{c-2}} \right)^{\frac{m-1}{32e^{4\gamma}}}$ . Then for  $s \geq \frac{3n}{4}$ ,

$$\begin{aligned} \frac{b_s}{b_{s-1}} &= \left( \frac{s}{n-s+1} \left( \frac{n-s+1}{n-s} \right)^{c-2} \right)^{\frac{m-1}{32e^{4\gamma}}} \\ &\geq \left( \frac{s}{n-s} \right)^{\frac{m-1}{32e^{4\gamma}}} \geq 9. \end{aligned}$$

We again bound another subset of the sum through a geometric series:

$$\begin{aligned} \sum_{s=\lceil 3n/4 \rceil}^{n-c+1} b_s &\leq \sum_{s=\lceil 3n/4 \rceil}^{n-c+1} \frac{b_{n-c+1}}{9^{n-c+1-s}} \\ &= \sum_{s=-\infty}^{n-c+1} \frac{b_{n-c+1}}{9^{n-c+1-s}} \\ &= \frac{9b_{n-c+1}}{8} \\ &= \frac{9}{8} \left( \frac{(c-1)!(n-c+1)!n^{c-1}}{n!(c-1)^{c-2}} \right)^{\frac{m-1}{32e^{4\gamma}}}. \end{aligned}$$

Similar to [6], we bound the last expression using Sterling's bound,  $\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n} \leq n! \leq en^{n+\frac{1}{2}}e^{-n}$ , so that  $\frac{9}{8} \left( \frac{(c-1)!(n-c+1)!n^{c-1}}{n!(c-1)^{c-2}} \right)^{\frac{m-1}{32e^{4\gamma}}} \leq \frac{9}{8} \left( \frac{e}{\sqrt{2\pi}} (c-1)^{1.5} \left( 1 - \frac{(c-1)}{n} \right)^{n-c+1.5} \right)^{\frac{m-1}{32e^{4\gamma}}}$ , which is maximized at  $c = 3$  for  $n \geq 19$ ,  $m \geq 8e^{4\gamma}$ , and  $c \leq \frac{n}{4}$ . Thus,

## 17:22 Differentially Private Aggregation via Imperfect Shuffling

$$\begin{aligned} \frac{9}{8} \left( \frac{e}{\sqrt{2\pi}} (c-1)^{1.5} \left( 1 - \frac{(c-1)}{n} \right)^{n-c+1.5} \right)^{\frac{m-1}{32e^{4\gamma}}} &\leq \frac{9}{8} \left( \frac{2e}{\sqrt{\pi}} \left( 1 - \frac{2}{n} \right)^{n-1.5} \right)^{\frac{m-1}{32e^{4\gamma}}} \\ &\leq \frac{9}{8} (1.27)^{\frac{m-1}{32e^{4\gamma}}}. \end{aligned}$$

**Upper bounding the middle terms in the summation.** It remains to upper bound the summation across  $s \in [\frac{n}{4e^{4\gamma}}, \frac{3n}{4}]$ . We have for  $\alpha = \frac{s}{n}$ ,  $b_s = \left( \frac{((1-\alpha)n)!(\alpha n)!}{(n-1)!(1-\alpha)^{c-2}} \right)^{\frac{m-1}{32e^{4\gamma}}}$ . By Sterling's bound, we have  $b_s \leq \left( \frac{e^2}{\sqrt{2\pi}} \sqrt{n} (1-\alpha)^{2.5-c+(1-\alpha)n} \alpha^{\alpha n + \frac{1}{2}} \right)^{\frac{m-1}{32e^{4\gamma}}} \leq \left( \frac{e^2 \sqrt{n}}{\sqrt{2\pi}} \alpha \right)^{\frac{m-1}{32e^{4\gamma}}}$ . Since there are at most  $n$  such terms  $b_s$ , then  $\sum_{s=k+1}^{\lceil 3n/4 \rceil - 1} b_s \leq n \left( \frac{e^2 \sqrt{n}}{\sqrt{2\pi}} \left( \frac{3}{4} \right)^{\frac{3n}{4}} \right)^{\frac{m-1}{32e^{4\gamma}}} \leq 2 \left( en \left( \frac{3}{4} \right)^{\frac{3n}{4}} \right)^{\frac{m-1}{32e^{4\gamma}}} \leq 2$ .

**Putting things together.** Combining the upper bounds across the three summations, we have

$$\begin{aligned} \sum_{s=1}^k \binom{n}{s}^{1-m} e^{2sm\gamma} \left( \frac{n^{c-1}}{(n-s)^{c-2}} \right)^{\frac{m-1}{32e^{4\gamma}}} + \sum_{s=k+1}^{n-c+1} \left( \frac{(n-s)!s!n^{c-1}}{n!(n-s)^{c-2}} \right)^{\frac{m-1}{32e^{4\gamma}}} \\ \leq \frac{26}{25} e^{m\gamma} e^{\frac{m-1}{32e^{4\gamma}}} + 2 + \frac{9}{8} (1.27)^{\frac{m-1}{32e^{4\gamma}}} \\ \leq 2e^{\frac{m-1}{32e^{4\gamma}}} \cdot e^{m\gamma} \leq 2e^{\frac{m-1}{32e^{4\gamma}}} \cdot e^{2\gamma(m-1)}. \end{aligned}$$

Therefore, we have  $p(n, c) \leq \frac{2^{c-1}}{c!} \left( \frac{e}{n} \right)^{\frac{(m-1)(c-1)}{32e^{4\gamma}}} \cdot e^{2\gamma(m-1)(c-1)}$ , as desired.  $\blacktriangleleft$