Privacy Amplification via Shuffling:
Unified, Simplified, and Tightened

Shaowei Wang
Guangzhou University
Guangzhou, China
wangsw@gzhu.edu.cn

Yun Peng
Guangzhou University
Guangzhou, China
yunpeng@gzhu.edu.cn

Jin Li
Guangzhou University
Guangzhou, China
lijin@gzhu.edu.cn

Zikai Wen
Virginia Tech
Virginia, U.S.A
zkwen@vt.edu

Zhipeng Li
Guangzhou University
Guangzhou, China
2006200002@e.gzhu.edu.cn

Shiyu Yu
Guangzhou University
Guangzhou, China
2112106255@e.gzhu.edu.cn

Di Wang
King Abdullah University of Science
and Technology
Thuwal, Saudi Arabia
di.wang@kaust.edu.sa

Wei Yang
University of Science and Technology
of China
Suzhou, China
qubit@ustc.edu.cn

ABSTRACT

The shuffle model of differential privacy provides promising privacy-utility balances in decentralized, privacy-preserving data analysis. However, the current analyses of privacy amplification via shuffling lack both tightness and generality. To address this issue, we propose the *variation-ratio reduction* as a comprehensive framework for privacy amplification in both single-message and multi-message shuffle protocols. It leverages two new parameterizations: the total variation bounds of local messages and the probability ratio bounds of blanket messages, to determine indistinguishability levels. Our theoretical results demonstrate that our framework provides tighter bounds, especially for local randomizers with extremal probability design, where our bounds are exactly tight. Additionally, variation-ratio reduction complements parallel composition in the shuffle model, yielding enhanced privacy accounting for popular sampling-based randomizers employed in statistical queries (e.g., range queries, marginal queries, and frequent itemset mining). Empirical findings demonstrate that our numerical amplification bounds surpass existing ones, conserving up to 30% of the budget for single-message protocols, 75% for multi-message ones, and a striking 75%-95% for parallel composition. Our bounds also result in a remarkably efficient $\tilde{O}(n)$ algorithm that numerically amplifies privacy in less than 10 seconds for $n = 10^8$ users.

1 INTRODUCTION

In our increasingly digital world, safeguarding data privacy has become eminent, especially in sensitive sectors like census [46], healthcare [67], and e-commerce [11]. As we gather data that contain references to identifiable personal information, the risk of privacy breaches emerges as a substantial threat to both individuals and organizations [60, 67]. In response to this growing concern, and in light of stringent data privacy regulations [16, 72], there is an urgent need for robust privacy protection measures, particularly in data analysis and learning. A promising answer lies in the concept of differential privacy (DP) [27], a mathematical framework ensuring data privacy while maintaining data utility.

DP traditionally operates in two models: the curator model and the local model, each serving different types of data analysis needs. The curator model, widely used for tasks like count queries [18, 19, 56, 57], data mining [15, 47, 59], and general SQL engines [49, 55], involves a trusted curator who applies privacy-preserving mechanisms before releasing data. The local model, on the other hand, offers privacy at the individual data source level, useful for queries like histogram query [9, 50], range query [23, 25, 68], marginal query [22, 90, 91], frequent itemset mining [82, 85] and machine learning [26, 71, 73, 74]. Despite their extensive applications in database, both models encounter difficulties in striking the balance between privacy and utility, particularly in decentralized settings where data control and processing are distributed among multiple participants.

The shuffle model of differential privacy [13, 30] emerges as a better solution that provides more effective balance between privacy protection and data utility in decentralized environments (e.g., federated analytics [8, 21, 32, 38, 39] and federated machine learning [35, 41, 42, 70]). In this model, taking the local ϵ_0-DP

mechanisms [77, 80, 83] as an instance, each user first applies a randomizer to transform their data into one or more messages. These messages are then sent to and mixed up by a "shuffler" before they being sent to the statistician (i.e., the server or the analyzer). The shuffler is the key player who obscures the origins of each piece of data. This process, known as *privacy amplification via shuffling*, significantly enhances the effectiveness of differential privacy by blocking the statistician from tracing data back to individual users.

Analyzing the level of privacy amplification is a critical step in maintaining a balanced trade-off between privacy protection and data utility. This analysis, however, faces significant challenges due to the expansive message space involved and the complex, non-linear nature of the shuffling operations on n messages. This complexity stands in contrast to simpler operations like summation in traditional DP approaches [14, 63].

To address these challenges, there has been considerable research effort. A notable research by Erlingsson *et al.* [30] demonstrated when each of the n shuffled messages is randomized using ϵ_0-LDP, they collectively maintain differential privacy at a level of $O(\epsilon_0\sqrt{\log(1/\delta)/n})$, with a high probability of $1 - \delta$, given that ϵ_0 is of order $O(1)$. For the more practical case of $\epsilon_0 = \Theta(1)$, another research [7] proposed viewing messages from other users as a "privacy blanket" that conceals each user's sensitive information. This approach showed that the shuffled messages preserve $(O(\min\{\epsilon_0, 1\}e^{\epsilon_0}\sqrt{\log(1/\delta)/n}), \delta)$-DP. Recent research [32, 33] introduced a "clone" concept to examine the indistinguishability offered by messages from other users, achieving an amplification bound of $(O(e^{\epsilon_0/2}\sqrt{\log(1/\delta)/n}), \delta)$-DP for $\epsilon_0 = \Theta(1)$.

While these analyses are significant, they have limitations in tightness and generality. The "privacy blanket" framework [7], for example, provides tighter parameters for specific ϵ_0-LDP randomizers. However, it is less precise in tail-bounding privacy loss variables, even when using their $\Omega(n)$-complexity numerical bounds. Similarly, the recent "clone" reduction [33] achieves nearly optimal amplification bounds for general ϵ-LDP randomizers. Yet, it falls short in offering precise amplification for common LDP mechanisms, such as Hadamard response [1], subset selection [77], and PrivUnit [14]. Consequently, these approaches do not fully assist practitioners in setting optimal privacy parameters to maximize data utility. Additionally, their results are only applicable to (single-message) shuffle protocols with LDP randomizers, which face intrinsic utility drawbacks [36]. Such drawbacks are not as pronounced in protocols that utilize the metric DP notion [17], or multi-message protocols where each user contributes multiple messages [8].

In the context of multi-message shuffle model, current works leverage random numerical shares [37, 38] or categorical shares [8, 58, 61] to collaboratively preserve privacy post-shuffling. Nevertheless, these protocols, including those in [21, 58, 61, 81], analyze DP guarantees on a case-by-case basis, offering only approximate/loose analytic bounds. This hampers their generalizability across different protocols and leads to undesirable practical utility.

1.1 Our Contributions

Recognizing the aforementioned limitations, our research introduces two advancements in analyzing the privacy amplification of differentially private data processing.

First, we designed a unified, tight, and scalable framework for analyzing privacy amplification within the shuffle model that supports simutaneously LDP randomizers, metric LDP randomizers, and various multi-message protocols. This framework, which we term the *variation-ratio* reduction, innovatively links DP levels to two novel yet intuitive parametrizations: the pairwise total variation and the one-direction probability ratio of local randomizers. It demonstrates exact tightness for an extensive family of randomizers that exhibit extremal properties. This includes a variety of state-of-the-art ϵ_0-LDP randomizers like local hash [83], Hadamard response [1], the PrivUnit mechanism [12], and several multi-message protocols [5, 21, 58, 61]. Moreover, our framework offers precise bounds with a computational complexity of $\tilde{O}(n)$, making it highly effective for large-scale data analyses, such as telemetry data involving millions or billions of users.

Second, we derived tight privacy amplification bounds for parallel local randomizers (i.e., sampling-based composite randomizers). These types of randomizers are widely deployed by differentially private protocols in user data collecting/queries, such as count queries [24, 69, 84], heavy hitter identification [10, 66, 86], frequent itemset mining [85], and machine learning [34, 48]. Within the variation-ratio framwork, our analysis reveals that the pairwise total variation bound of a parallel randomizer does not exceed the expected bound of its individual base randomizers. This insight forms the basis of our *advanced parallel composition theorem* in the shuffle model, enabling much tighter amplification bounds compared to straightforward approaches.

1.2 Organization

The remaining paper is organized as follows. Section 2 reviews related works. Section 3 provides background knowledge. Section 4 presents main theoretical upper bounding results, the corresponding proof sketch, and an efficient numerical bounding algorithm. Section 5 establishes amplification lower bounds. Section 6 improves bounding results for parallel randomizers. Section 7 shows experimental results. Finally, Section 8 concludes the paper.

2 RELATED WORK

Privacy amplification of LDP randomizers. The seminal work [30] utilizes the privacy amplification via subsampling [6, 52] to analyze the shuffling amplification, and shows n shuffled messages satisfies $(\epsilon_0\sqrt{144\log(1/\delta)/n}, \delta)$-DP. Latterly, the privacy blanket [7] proposes extracting an input-independent part from the output distribution, to work as a "blanket" to amplify privacy. For general LDP randomizers, the [7] shows the input-independent part is sampled by each user with probability at least $e^{-\epsilon_0}$ (they term it as total variation similarity). For specific LDP randomizers (e.g., Laplace [29] and generalized randomized response [88]), the total variation similarity can be larger. Recently, the works [32] and [33] decompose output distributions into mixture distributions with 3 options, and interprets messages from other users as clones of victim user. This clone reduction is near-optimal w.r.t. the dependence on ϵ_0 for general LDP randomizers. Meanwhile, in order to derive tighter amplification bounds for specific randomizers, it needs to explicitly construct mixture distributions and determine the clone probabilities, which however is intractable for commonly

deployed randomizers (e.g., Laplace mechanism, subset selection mechanism [77, 92], PrivUnit mechanism [14], and sampling-based composite randomizers). As comparision, our framework using a new parametrization (i.e., pairwise total variation) of randomizers to implicitly derive tighter clone probabilities. Besides, the clone reductions in research [32, 33] are limited to LDP randomizers, while our framework generalizes to metric-based LDP randomizers and multi-message protocols. Even for LDP randomizers in the shuffle model, our asymptotic amplification bounds are provably tighter than existing results (see Table 1).

Privacy amplification of metric-based LDP randomizers. The metric-based LDP [17] permits a flexible indistinguishable level between any pair of inputs $a, b \in \mathbb{X}$ (denoted as $d_{\mathbb{X}}(a, b)$), and is widely used for user data with a large domain range such as location data [3] and unstructed data [94]. Previous work [78] initiated the study of metric DP in the shuffle model and provided several amplification bounds, which depend on $\max_{c \in \mathbb{X}} e^{d_{\mathbb{X}}(a,c)} + e^{d_{\mathbb{X}}(b,c)}$. In our work, we apply the proposed framework to metric LDP randomizers and improve the dependence to $(1 - e^{-d_{\mathbb{X}}(a,b)}) \cdot \max_{c \in \mathbb{X}} \max[e^{d_{\mathbb{X}}(a,c)}, e^{d_{\mathbb{X}}(b,c)}]$ (a lower value indicates stronger amplification effects, see Section 4.3 for detail).

Privacy amplification of multi-message randomizers. For multi-message protocols in the shuffle model (e.g., [21, 58, 61]), the local randomizer may neither satisfy LDP nor metric-based LDP. Hence, designated analyses of privacy guarantees are often conducted. Our proposed framework yields much tighter privacy guarantees and reduces the budget by 70%-85% compared to these designated analyses (see detail in Section 7).

Privacy amplification via shuffling under composition. To analyze accumulated privacy loss under sequential composition for the shuffle model [33, 40, 41, 43, 53], it is common to use analytic/numerical tools from the centralized model of differential privacy (e.g., strong composition theorem [28], Rényi differential privacy [62], and Fourier accountant [54]). The [43] analyzes Rényi differential privacy of shuffled messages, [40] further composes privacy for the shuffle model with subsamplings. To the best of our knowledge, we are the first to provide tight privacy amplification results for prevalent parallel composition (see detail in Section 6).

3 PRELIMINARIES

Definition 3.1 (Hockey-stick divergence). The Hockey-stick divergence (with parameter e^{ϵ}) between two random variables P and Q is defined by:

$$D_{e^{\epsilon}}(P\|Q) = \int \max\{0, P(x) - e^{\epsilon}Q(x)\}\mathrm{d}x,$$

where we use the notation P and Q to refer to both the random variables and their probability density functions.

Two variables P and Q are (ϵ, δ)-indistinguishable iff

$$\max\{D_{e^{\epsilon}}(P\|Q), D_{e^{\epsilon}}(Q\|P)\} \leq \delta.$$

If two datasets have the same size and differ only by the data of a single individual, they are referred to as neighboring datasets. Differential privacy ensures that the divergence of query outputs on neighboring datasets is constrained (see Definition 3.2). Similarly, in the local setting where a single individual's data is taken as input,

we present the definition of local (ϵ, δ)-DP in Definition 3.3. The notation ϵ-LDP is used when the failure probability $\delta = 0$.

Definition 3.2 (Differential privacy [29]). An algorithm $\mathcal{R} : \mathbb{X}^n \mapsto \mathbb{Z}$ satisfies (ϵ, δ)-differential privacy iff for all neighboring datasets $X, X' \in \mathbb{X}^n$, $\mathcal{R}(X)$ and $\mathcal{R}(X')$ are (ϵ, δ)-indistinguishable.

Definition 3.3 (Local differential privacy [52]). An algorithm $\mathcal{R} : \mathbb{X} \mapsto \mathbb{Y}$ satisfies local (ϵ, δ)-differential privacy iff for all $x, x' \in \mathbb{X}$, $\mathcal{R}(x)$ and $\mathcal{R}(x')$ are (ϵ, δ)-indistinguishable.

Metric differential privacy. A distance measure $d_{\mathbb{X}} : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}$ over $\mathbb{X}$ characterizes the distances between any pair of elements $a, b \in \mathbb{X}$. For this measure to be deemed a metric, it must adhere to the subsequent properties:

i. *Self-identity:* For any $a \in \mathbb{X}$, $d_{\mathbb{X}}(a, a) = 0$.
ii. *Positivity:* For any $a, b \in \mathbb{X}$ that $a \neq b$, $d_{\mathbb{X}}(a, b) > 0$.
iii. *Symmetry:* For any $a, b \in \mathbb{X}$, $d_{\mathbb{X}}(a, b) = d_{\mathbb{X}}(b, a)$.
iv. *Triangle inequality:* For any $a, b, c \in \mathbb{X}$, $d_{\mathbb{X}}(a, b) + d_{\mathbb{X}}(b, c) \geq d_{\mathbb{X}}(a, c)$.

The concept of metric differential privacy expands upon the traditional differential privacy, enabling the calibration of indistinguishability constraints to element dissimilarity.

Definition 3.4 (Metric $(d_{\mathbb{X}}, \delta)$-differential privacy [17]). A randomized mechanism $\mathcal{R}$ satisfies $(d_{\mathbb{X}}, \delta)$-differential privacy iff for any neighboring datasets $X, X' \in \mathbb{X}^n$ (i.e., $x_i = a$ in X and $x_i = b$ in X'), $\mathcal{R}(X)$ and $\mathcal{R}(X')$ are $(d_{\mathbb{X}}(a, b), \delta)$-indistinguishable.

In the local setting, each user sanitizes their data on their own locally. Substituting the dataset S in Definition 3.4 with an individual data item, the definition of metric LDP is as follows:

Definition 3.5 (Local metric $d_{\mathbb{X}}$-differential privacy [2, 3]). Let $\mathbb{D}_{\mathcal{R}}$ denote the output domain, a randomized mechanism $\mathcal{R}$ satisfies local metric $d_{\mathbb{X}}$-privacy iff for any $a, b \in \mathcal{X}$, $\mathcal{R}(a)$ and $\mathcal{R}(b)$ are $(d_{\mathbb{X}}(a, b), 0)$-indistinguishable.

3.1 The Shuffle Model of Differential Privacy

We use $[n]$ to denote $\{1, ..., n\}$ and $[i : j]$ to denote $\{i, i + 1, ..., j\}$. Follow conventions in the shuffle model based on randomize-then-shuffle [7, 20], we define a single-message protocol $\mathcal{P}$ to be a list of algorithms $\mathcal{P} = (\{\mathcal{R}_i\}_{i \in [n]}, \mathcal{A})$, where $\mathcal{R}_i : \mathbb{X} \to \mathbb{Y}$ is the local randomizer of user i, and $\mathcal{A} : \mathbb{Y}^n \to \mathbb{Z}$ the analyzer in the data collector's side. The overall protocol implements a mechanism $\mathcal{P} : \mathbb{X}^n \to \mathbb{Z}$ as follows. Each user i holds a data record x_i and a local randomizer $\mathcal{R}_i$, then computes a message $y_i = \mathcal{R}_i(x_i)$. The messages $y_1, ..., y_n$ are then shuffled and submitted to the analyzer. We write $\mathcal{S}(y_1, \ldots, y_n)$ to denote the random shuffling step, where $\mathcal{S} : \mathbb{Y}^n \to \mathbb{Y}^n$ is a *shuffler* that applies a uniform-random permutation to its inputs. In summary, the output of $\mathcal{P}(x_1, \ldots, x_n)$ is denoted by $\mathcal{A} \circ \mathcal{S} \circ \mathcal{R}_{[n]}(X) = \mathcal{A}(\mathcal{S}(\mathcal{R}_1(x_1), \ldots, \mathcal{R}_n(x_n)))$.

The shuffle model assumed that all parties involved in the protocol follow it faithfully and there is no collusion between them. From a privacy perspective, the goal is to ensure the differential privacy of the output $\mathcal{P}(x_1, \ldots, x_n)$ for any analyzer $\mathcal{A}$. By leveraging the post-processing property of the Hockey-stick divergence, it suffices to ensure that the shuffled messages $\mathcal{S} \circ \mathcal{R}_{[n]}(X) = \mathcal{S}(\mathcal{R}_1(x_1), \ldots, \mathcal{R}_n(x_n))$ are differentially private. We formally define differential privacy in the shuffle model in Definition 3.6.

Definition 3.6 (Differential privacy in the shuffle model). A protocol $\mathcal{P} = (\{\mathcal{R}_i\}_{i\in[n]}, \mathcal{A})$ satisfies (ϵ, δ)-differential privacy in the shuffle model iff for all neighboring datasets $X, X' \in \mathbb{X}^n$, the $\mathcal{S} \circ \mathcal{R}_{[n]}(X)$ and $\mathcal{S} \circ \mathcal{R}_{[n]}(X')$ are (ϵ, δ)-indistinguishable.

In multi-message shuffle protocols, which offer greater utility potential than single-message counterparts [36], each user can submit multiple messages. These messages from all users are then randomly shuffled. The differential privacy applied in this context is determined w.r.t. a single alteration in the input dataset, analogous to the approach in single-message protocols.

4 THE VARIATION-RATIO FRAMEWORK

In this section, we introduce our variation-ratio reduction framework for privacy amplification via shuffling. Our approach leverages the mixture decomposition of local randomizers, as introduced in the privacy blanket [7] and clone reduction [32, 33] literature. The mixture decomposition of local randomizers reveal two important facts: (1) randomized and anonymously shuffled messages from other users can mimic as the message from a particular (victim) user with some probability, which are termed as blanket messages in [7] and clones in [32, 33] and can thus amplify privacy; (2) for the local randomizer of the victim user that pertain to some properties (e.g., $(\epsilon, 0)$-LDP), the probability distribution of outputs given arbitrary inputs share some resemblance. The key to analyzing tighter shuffle privacy amplification would be characterizing more concise mimic probability from other users and the probability resemblance of the victim's outputs. Compared to existing results, our framework achieves a highly precise mixture decomposition by utilizing both $(0, \delta)$-LDP and $(\epsilon_0, 0)$-LDP properties of local randomizers to capture the the probability resemblance of victim's outputs (over differed inputs), and introduces the probability ratio as simple yet effective indicates to the mimic power from other users.

To establish our approach, without loss of generality, we consider two neighboring datasets X and X' that differ only in the first user's data (i.e. the first user is considered as the victim in DP), i.e., $X = \{x_1 = x_1^0, x_2, \ldots, x_n\}$ and $X' = \{x_1 = x_1^1, x_2, \ldots, x_n\}$, where $x_1^0, x_1^1, x_2, \ldots, x_n \in \mathbb{X}$. We define the following properties on (independent) local randomizers $\{\mathcal{R}_i\}_{i\in[n]}$ with some parameters $p > 1, \beta \in [0, \frac{p-1}{p+1}]$, and $q \geq 1$:

I. (p, β)-*variation property*: we say that the (p, β)-variation property holds if $D_p(\mathcal{R}_1(x_1^0)\|\mathcal{R}_1(x_1^1)) = 0$ and $D_{e^0}(\mathcal{R}_1(x_1^0)\| \mathcal{R}_1(x_1^1)) \leq \beta$ for all possible $x_1^0, x_1^1 \in \mathbb{X}$.

II. q-*ratio property*: we say that the q-ratio property holds if $D_q(\mathcal{R}_1(x_1)\|\mathcal{R}_i(x_i)) = 0$ for all possible $x_1, \ldots, x_n \in \mathbb{X}$ and all $\{\mathcal{R}_i\}_{i\in[2:n]}$.

The parameter β corresponds to the pairwise total variation distance and serves as an indicator of the degree to which $\mathcal{R}_1$ satisfies $(0, \beta)$-LDP. Conversely, the parameter p relates to the divergence ratio and serves as an indicator of the $(\log p, 0)$-LDP property satisfied by $\mathcal{R}_1$. Meanwhile, the parameter q captures the ability of $\mathcal{R}_i(x_i)$ to mimic $\mathcal{R}_1(x_1)$. For any randomizer $\mathcal{R}_1$, there exists a dominating pair of distributions [95, Proposition 8], such that the divergence upper bound (including the (p, β) parameters) can be directly derived upon. Specificlly, in case the randomizer $\mathcal{R}_1$ satisfy ϵ_0-LDP, we always have $p \leq \exp(\epsilon_0)$ and $\beta \leq \frac{e^{\epsilon_0}-1}{e^{\epsilon_0}+1}$ [51]. Most commonly-used ϵ_0-LDP randomizers have a lower total variation bound β than the worst-case $\frac{e^{\epsilon_0}-1}{e^{\epsilon_0}+1}$ (see Tables 2), thus permits a tighter mixture decomposition of their shuffled outputs.

4.1 Main Results

Our objective is to establish an upper bound on the divergence between the shuffled messages resulting from two independent protocol runs: $\mathcal{S}(\mathcal{R}_1(x_1^0), \ldots, \mathcal{R}_n(x_n))$ and $\mathcal{S}(\mathcal{R}_1(x_1^1), \ldots, \mathcal{R}_n(x_n))$, where the local randomizers $\{\mathcal{R}_i\}_{i\in[n]}$ satisfy the properties mentioned earlier. We accomplish this through an implicit analysis of the mixture decomposition of these local randomizers, then utilizing the data processing inequality to obtain a dominating pair of binomial counts. Subsequently, we represent the Hockey-stick divergence as an expectation of cumulative probabilities over a binomial variable c (see Theorem 4.1), exploiting the monotone of the integral in Hockey-stick divergence. The notation $\mathrm{CDF}_{c,1/2}[c_1, c_2]$ represents the cumulative probability $\sum_{i\in[c_1,c_2]} \binom{c}{i}/2^c$, which can be quickly computed using two calls to the regularized incomplete beta function [65]. Therefore, Theorem 4.1 directly implies an efficient algorithm for numerically compute Hockey-stick divergence within $\tilde{O}(n)$ complexity (see Section 4.4 for detail).

THEOREM 4.1 (DIVERGENCE UPPER BOUND). *For $p > 1, \beta \in [0, \frac{p-1}{p+1}]$, $q \geq 1$, if randomizers $\{\mathcal{R}_i\}_{i\in[n]}$ satisfy the (p, β)-variation property and the q-ratio property, then for any $x_1^0, x_1^1, x_2, \ldots, x_n \in \mathbb{X}$:*

$$D_{e^\epsilon}(\mathcal{S}(\mathcal{R}_1(x_1^0), \ldots, \mathcal{R}_n(x_n)) \| \mathcal{S}(\mathcal{R}_1(x_1^1), \ldots, \mathcal{R}_n(x_n)))$$

$$\leq \mathop{\mathbb{E}}_{c\sim Binom(n-1,2r)} \Big[(p - e^\epsilon)\alpha \cdot \mathop{\mathrm{CDF}}_{c,1/2}\big[\lceil low_{c+1} - 1\rceil, c\big]$$

$$+ (1 - pe^\epsilon)\alpha \cdot \mathop{\mathrm{CDF}}_{c,1/2}\big[\lceil low_{c+1}\rceil, c\big]$$

$$+ (1 - e^\epsilon)(1 - \alpha - p\alpha) \cdot \mathop{\mathrm{CDF}}_{c,1/2}\big[\lceil low_c\rceil, c\big]\Big],$$

where $\alpha = \frac{\beta}{p-1}$, $r = \frac{\alpha p}{q}$, $low_c = \frac{(e^\epsilon p - 1)\alpha c + (e^\epsilon - 1)(1 - \alpha - \alpha p)\cdot \frac{(n-c)r}{1-2r}}{\alpha(e^\epsilon + 1)(p-1)}$.

Asymptotically, when n is sufficiently large, we present two analytical amplification bounds: one that is sophisticated, and the other that is more concise but less tight. Specifically, we derive the former bound in Theorem 4.2. To achieve this, we tail bound the variable c (i.e. number of clones in the former theorem), and exploit the fact that the statistical divergence monotonically non-increases with c. The resulting bound is satisfactorily tight for a wide range of variation-ratio parameters (see Section 7.2 for comparison with numerical upper bounds). To tail bound c, we use multiplicative Chernoff bounds and Hoeffding's inequality, which induce tighter bounds for small clone probability $2r$ (in local $d_\mathbb{X}$-DP randomizers) and for large $2r$ (in multi-message protocols), respectively.

THEOREM 4.2 (ANALYTIC PRIVACY AMPLIFICATION BOUNDS). *For $p > 1, \beta \in [0, \frac{p-1}{p+1}]$, and $q \geq 1$, the $P_{p,\beta}^q$ and $Q_{p,\beta}^q$ are (ϵ, δ)-indistinguishable with*

$$\epsilon = \log\Bigg(1 + \frac{\beta(2\sqrt{\Omega\log(\frac{4}{\delta})/2} + 1) + \beta(\Omega/2 - \sqrt{\Omega\log(\frac{4}{\delta})/2})}{\alpha\Omega}$$

$$+ \frac{(1 - \alpha - \alpha p)(n - 1 - \Omega)r}{(1 - 2r)\alpha\Omega}\Bigg)$$

when $(p+1)\alpha/2 - (1 - \alpha - \alpha p)r/(1 - 2r) \geq 0$ and $\Omega = 2r(n-1) - \sqrt{\min(6r, \frac{1}{2})(n-1)\log(\frac{4}{\delta})} \geq \frac{2p(\beta+1+(\beta-1)p)(n-1)+\beta}{q+p(\beta-1+(\beta+1)p)-pq}$, where $\alpha = \beta/(p-1)$ and $r = p\alpha/q$.

We also derive a succinct formula in Theorem 4.3 by further tail bounding the number of clones c.

Theorem 4.3 (Asymptotic privacy amplification bounds). *For $p > 1, \beta \in [0, \frac{p-1}{p+1}], q \geq 1$, when $n \geq \frac{8\log(2/\delta)(p-1)q}{\beta p}$, the $P^q_{p,\beta}$ and $Q^q_{p,\beta}$ are (ϵ, δ)-indistinguishable with*

$$\epsilon = \log\left(1 + \frac{\beta}{(1-v)(1+p)\beta/(p-1)+v}\left(\sqrt{\frac{32\log(\frac{4}{\delta})}{r(n-1)}} + \frac{4}{rn}\right)\right),$$

where $r = \frac{p\beta}{(p-1)q}$ and $v = \max\{0, \frac{4}{9}\frac{1-3r}{1-2r}\}$.

When n is sufficiently large and a mild condition $r = \frac{p\beta}{(p-1)q} \leq 1/4$ holds, we have the v in Theorem 4.3 no less than a constant value $2/9$, thus the privacy can be amplified to:

$$\tilde{O}\left(\sqrt{\beta(p-1)q/(pn)}\right).$$

Therefore, when β and q are relatively small, there are strong privacy amplification effects. This corresponds to qualitative interpretation about β and q: the parameter β indicates the total variation distance between $R_1(x_1^0)$ and $R_1(x_1^1)$, so when β is relatively low, the statistical distance of shuffled messages would also be low; a lower q indicates that messages from other users (i.e. users $[2 : n]$) can better mimic $R_1(x_1^0)$ and $R_1(x_1^1)$, thus reduces the chance an adversary can disguise between x_1^0 and x_1^1 from shuffled messages.

Specifically, for randomizers satisfying ϵ_0-LDP, we have $q = p = \exp(\epsilon_0)$, thus the asymptotic amplified DP level becomes:

$$\tilde{O}\left(\sqrt{\beta(\exp(\epsilon_0) - 1)/n}\right).$$

In Table 1, we compare our asymptotic bound with existing asymptotic bounds found in research [7, 30, 32, 33], under the practical setting where $\epsilon_0 = \Theta(1)$. Since the total variation bound β is often significantly lower than the worst-case value of $\frac{e^{\epsilon_0}-1}{e^{\epsilon_0}+1}$, our amplification bound is provably tighter than previously known bounds.

Table 1: Comparison of privacy amplification bound on shuffled ϵ_0-LDP messages when $\epsilon_0 = \Theta(1)$ and $n \geq \tilde{\Omega}(\exp(\epsilon_0))$.

Method	Asymptotic Amplification Bound
EFMRTT19 [30]	$O(\sqrt{\exp(3\epsilon_0)\log(1/\delta)/n})$
privacy blanket [7]	$O(\sqrt{\exp(2\epsilon_0)\log(1/\delta)/n})$
clone [32]	$O\left(\frac{\exp(\epsilon_0)-1}{\exp(\epsilon_0)+1}\sqrt{\frac{\exp(\epsilon_0)\log(1/\delta)}{n}}\right)$
stronger clone [33]	$O\left(\sqrt{\frac{(\exp(\epsilon_0)-1)^2\log(1/\delta)}{n(\exp(\epsilon_0)+1)}}\right)$
variation ratio [**this work**]	$O\left(\sqrt{\frac{\beta(\exp(\epsilon_0)-1)\log(1/\delta)}{n}}\right)$

Discussion on differences with stronger clone reduction. For ϵ_0-LDP randomizers where $p \equiv q = \exp(\epsilon_0)$, our intermediate result presented in Theorem 4.7 yields a divergence upper bound analogous to the one found in the state-of-the-art stronger clone by [33]. Our framework offers three primary advantages: (i) Our

bound is tighter than [33] as long as the newly introduced total variation parameter β is not the worst-case $(p-1)/(p+1)$. Notably, commonly-used randomizers exhibit much lower β values, and computing β is straightforward. (ii) While [33] handles only LDP randomizers that $q \equiv p$, our reduction is versatile, accommodating also cases where $p < q$ (e.g., for metric local randomizers) and $p > q$ (e.g., for multi-message randomizers). (iii) Our method entails a $\tilde{O}(n)$-time algorithm for numerically determining the Hockey-stick divergence, offering a much more efficient solution compared to prevailing algorithms [32, 53], being about 10 times faster than their documented execution times.

4.2 Proof Sketch

Our proof for Theorem 4.1 follows three steps: (i) utilizing the (p, β)-variation and q-ratio properties to obtain mixture decomposition of local randomizers; (ii) motivated by the clone reduction [32, 33], we succinctly reduce the mixture decomposition as several binomial counts, which serve as the dominating pair of distribution for privacy analyses [95]; (iii) delving into the Hockey-stick divergence of the dominating binomial counts, we utilize the monotonicness of the Hockey-stick divergence to derive the final results of expectation on cumulative probabilities.

We first consider fixed x_1^0 and x_1^1, and show that $R_1(x_1^0), R_1(x_1^1)$, and $R_i(x_i)$ can be interpreted as mixture distributions (as elaborated in Lemma 4.4) with parameters related to p, q, and the total variation $\beta' = D_1(R_1(x_1^0)\|R_1(x_1^1))$. The core idea is to concentrate on elements where the distributions $R_1(x_1^0)$ and $R_1(x_1^1)$ differ, and bound the total differing probability using parameters p and β. For messages from other users, the parameter q is used to compute their likelihood of appearing as a message from user 1.

Lemma 4.4 (Mixture decompositions). *Given $x_1^0, x_1^1, ..., x_n \in \mathbb{X}$, if algorithms $\{R_i\}_{i \in [n]}$ satisfy the (p, β')-variation property and the q-ratio property with some $p > 1, q \geq 1$ and $\beta' = D_1(R_1(x_1^0)\|R_1(x_1^1))$, then there exists distributions $Q_1^0, Q_1^1, Q_1, Q_2, ..., Q_n$ such that*

$$R_1(x_1^0) = p\alpha Q_1^0 + \alpha Q_1^1 + (1 - \alpha - p\alpha)Q_1 \quad (1)$$

$$R_1(x_1^1) = \alpha Q_1^0 + p\alpha Q_1^1 + (1 - \alpha - p\alpha)Q_1 \quad (2)$$

$$\forall i \in [2, n], R_i(x_i) = rQ_1^0 + rQ_1^1 + (1 - 2r)Q_i \quad (3)$$

where $\alpha = \frac{\beta'}{p-1}$ and $r = \frac{\alpha p}{q}$.

Next, we generalize the clone reduction technique [33] by leveraging the data processing inequality to relate mixture parameters to the level of divergence (as demonstrated in Lemma 4.5). Specifically, this generalized reduction lemma implies that the statistical distance between shuffled messages

$$D(S(R_1(x_1^0), ..., R_n(x_n)) \| S(R_1(x_1^1), ..., R_n(x_n)))$$

is bounded by the one between binomial counts: $D(P^q_{p,\beta'}\|Q^q_{p,\beta'})$.

Lemma 4.5 (Divergence reduction via mixture). *Given any $n + 1$ inputs $x_1^0, x_1^1, x_2, ..., x_n \in \mathbb{X}$, consider algorithms $\{R_i\}_{i \in [n]}$ such*

that the output domain is finite and

$$\mathcal{R}_1(x_1^0) = p\alpha Q_1^0 + \alpha Q_1^1 + (1 - \alpha - p\alpha)Q_1,$$

$$\mathcal{R}_1(x_1^1) = \alpha Q_1^0 + p\alpha Q_1^1 + (1 - \alpha - p\alpha)Q_1,$$

$$\forall i \in [2, n], \ \mathcal{R}_i(x_i) = rQ_1^0 + rQ_1^1 + (1 - 2r)Q_i$$

holds for some $p \geq 1, q > 1, \alpha = \beta'/(p-1) \in [0, 1/(p+1)], r \in [0, 1/2]$ and some probability distributions $Q_1^0, Q_1^1, Q_1, Q_2, ..., Q_n$. Let $C \sim Binom(n-1, 2r)$, $A \sim Binom(C, 1/2)$, and $\Delta_1 = Bernoulli(p\alpha)$ and $\Delta_2 = Bernoulli(1 - \Delta_1, \alpha/(1 - p\alpha))$; let $P_{p,\beta'}^q = (A + \Delta_1, C - A + \Delta_2)$ and $Q_{p,\beta'}^q = (A + \Delta_2, C - A + \Delta_1)$. Then for any distance measure D that satisfies the data processing inequality,

$$D(\mathcal{S}(\mathcal{R}_1(x_1^0), .., \mathcal{R}_n(x_n)) \| \mathcal{S}(\mathcal{R}_1(x_1^1), .., \mathcal{R}_n(x_n))) \leq D(P_{p,\beta'}^q \| Q_{p,\beta'}^q).$$

Then, we utilize the fact that the statistical distance $D(P_{p,\beta'}^q \| Q_{p,\beta'}^q)$ is monotonically non-decreasing with β'. For any arbitrary x_1^0 and x_1^1, the total variation is upper bounded by β. Thus, we conclude that for all possible neighboring datasets, the statistical distance between the shuffled messages is upper bounded by $D(P_{p,\beta}^q \| Q_{p,\beta}^q)$.

LEMMA 4.6 (NON-DECREASING OF DIVERGENCE). *For any $p > 1, q \geq 1$ and $\beta, \beta' \in [0, \frac{p-1}{p+1}]$, if $\beta > \beta'$, then for any distance measure D that satisfies data processing inequality,*

$$D(P_{p,\beta}^q \| Q_{p,\beta}^q) \geq D(P_{p,\beta'}^q \| Q_{p,\beta'}^q).$$

Use the fact that β' is upper bounded by β, the results of previous steps can be summarized in in Theorem 4.7. It is applicable to any divergence measures that satisfy the data-processing inequality, such as Rényi divergences [62].

THEOREM 4.7 (VARIATION-RATIO REDUCTION). *For $p > 1, \beta \in [0, \frac{p-1}{p+1}], q \geq 1$, let $C \sim Binom(n-1, \frac{2\beta p}{(p-1)q})$, $A \sim Binom(C, 1/2)$ and $\Delta_1 \sim Bernoulli(\frac{\beta p}{p-1})$ and $\Delta_2 \sim Bernoulli(1 - \Delta_1, \frac{\beta}{p-1-\beta p})$; let $P_{p,\beta}^q$ denote $(A + \Delta_1, C - A + \Delta_2)$ and $Q_{p,\beta}^q$ denote $(A + \Delta_2, C - A + \Delta_1)$. If randomizers $\{\mathcal{R}_i\}_{i \in [n]}$ satisfy the (p, β)-variation property and the q-ratio property, then for any $x_1^0, x_1^1, x_2, ..., x_n \in \mathbb{X}$ and any measurement D satisfying the data-processing inequality:*

$$D(\mathcal{S}(\mathcal{R}_1(x_1^0), ..., \mathcal{R}_n(x_n)) \| \mathcal{S}(\mathcal{R}_1(x_1^1), ..., \mathcal{R}_n(x_n))) \leq D(P_{p,\beta}^q \| Q_{p,\beta}^q).$$

THEOREM 4.8 (DIVERGENCE BOUND AS AN EXPECTATION). *For $p > 1, \beta \in [0, \frac{p-1}{p+1}], q \geq 1$, let $\alpha = \beta/(p-1)$ and $r = \alpha p/q$, then for any $\epsilon \in \mathbb{R}$, the following two equations hold:*

$$D_{e^\epsilon}(P_{p,\beta}^q \| Q_{p,\beta}^q) = \mathop{\mathbb{E}}_{c \sim Binom(n-1, 2r)} \left[(p - e^\epsilon)\alpha \cdot \mathop{CDF}_{c,1/2}[\lceil low_{c+1} - 1 \rceil, c] \right.$$
$$+ (1 - pe^\epsilon)\alpha \cdot \mathop{CDF}_{c,1/2}[\lceil low_{c+1} \rceil, c]$$
$$\left. + (1 - e^\epsilon)(1 - \alpha - p\alpha) \cdot \mathop{CDF}_{c,1/2}[\lceil low_c \rceil, c] \right],$$

where $low_c = \frac{(e^{\epsilon'} p - 1)\alpha c + (e^{\epsilon'} - 1)(1 - \alpha - \alpha p)(n-c) \cdot \frac{r}{1-2r}}{\alpha(e^{\epsilon'} + 1)(p-1)}$.

Finally, we show an efficient way to computing (the upper bound of) $D_{e^{\epsilon'}}(P_{p,\beta}^q \| Q_{p,\beta}^q)$. A straight-forward approach is enumerating the output space $(a, b) \in [0, n]^2$ using the definition of the divergence. However, this method has a high computational cost of $O(n^2)$ and may suffer from numerical underflow issues. To avoid

these issues, we exploit the monotonicity of the probability ratio $\frac{\mathbb{P}[P_{p,\beta}^q = (a,b)]}{\mathbb{P}[Q_{p,\beta}^q = (a,b)]}$ with respect to a when $a + b$ is fixed. Thus, the maximum operation in the Hockey-stick divergence can be safely removed by carefully tracking the range of a, and the overall divergence can be expressed as an expectation over $c \sim Binom(n-1, 2r)$ that follows binomial distribution (see Theorem 4.8). To track the range of a, we introduce the notation $CDF_{c,1/2}[c_1, c_2]$ to represent the cumulative probability $\sum_{i \in [c_1, c_2]} \binom{c}{i}/2^c$, which can be computed using two calls to the regularized incomplete beta function [65]. As a result, the computational complexity is reduced to $\tilde{O}(n)$.

4.3 Amplification Parameters of Randomizers

To demonstrate the broad applicability of our proposed framework for privacy amplification analysis in the shuffle model, we provide amplification parameters for a wide range of local randomizers, including prevalent ϵ_0-LDP randomizers, local metric $d_{\mathbb{X}}$-DP randomizers, and multi-message randomizers.

LDP randomizers. We summarize the variation-ratio parameters of ϵ-LDP randomizers in Table 2. The worst-case total variation bound $\beta = \frac{e^\epsilon - 1}{e^\epsilon + 1}$ for general mechanisms is derived from [51], which proves that randomized response maximizes it. This worst-case bound is equivalent to the asymptotically optimal stronger clone reduction in research [33]. For commonly used mechanisms, such as mean estimation mechanisms [26, 29, 64, 75], distribution estimation mechanisms [1, 26, 31, 50, 77, 83, 92], and mechanisms for complicated data [45, 66, 76, 79], we exploit their specific structures to derive tighter total variation bounds, which can lead to stronger amplification effects (see Lemma 4.6 for qualitative analyses and Section 7 for numerical results).

Local metric DP radomizers. Regarding local metric DP mechanisms, we study the indistinguishability level between shuffled messages generated by a metric locally $d_{\mathbb{X}}$-private mechanism $\mathcal{R}$, where $D_{\exp(d_{\mathbb{X}}(x,x'))}(\mathcal{R}(x) \| \mathcal{R}(x')) = 0$ holds for all $x, x' \in \mathbb{X}$. Specifically, we aim to analyze the indistinguishable level between $\mathcal{S}(\mathcal{R}(x_1^0), ..., \mathcal{R}(x_n))$ and $\mathcal{S}(\mathcal{R}(x_1^1), ..., \mathcal{R}(x_n)))$, which is also captured by Theorem 4.7 with $p \leq \exp(d_{01})$ and $q \leq \exp(d_{max})$. Here, d_{01} denotes the local indistinguishable level $d_{\mathbb{X}}(x_1^0, x_1^1)$, and d_{max} denotes the maximum indistinguishable level

$$\max_{x \in \mathbb{X}} \max\{d_{\mathbb{X}}(x, x_1^0), d_{\mathbb{X}}(x, x_1^1)\}$$

w.r.t. x_1^0 and x_1^1. We summarize the variation-ratio parameters for $\mathcal{S}(\mathcal{R}(x_1^0), ..., \mathcal{R}(x_n))$ and $\mathcal{S}(\mathcal{R}(x_1^1), ..., \mathcal{R}(x_n)))$ in Table 3. For general local $d_{\mathbb{X}}$-DP mechanisms, the amplification parameters $p = e^{d_{01}}, \beta = \frac{e^{d_{01}} - 1}{e^{d_{01}} + 1}$ come directly from d_{01}-LDP properties of $\mathcal{R}$ on $x_1 \in \{x_1^0, x_1^1\}$, the parameter $q = e^{d_{max}}$ comes from the metric privacy constraints that $\mathcal{R}(x_i)$ and $\mathcal{R}(x_1)$ are $(d_{\mathbb{X}}(x_i, x_1))$-indistinguishable and $d_{\mathbb{X}}(x_i, x_1) \leq d_{max}$. Compared with the amplification upper bound for general mechanisms proved by [78], where the clone probability $2r = 2/(\max_{x \in \mathbb{X}} e^{d_{\mathbb{X}}(x, x_1^0)} + e^{d_{\mathbb{X}}(x, x_1^1)})$, our clone probability $2\frac{\beta p}{(p-1)q} = 2/(e^{d_{max}} + e^{d_{max} - d_{01}})$ is at least not smaller than theirs (due to the triangle inequality property of $d_{\mathbb{X}}$). For Laplace mechanism with ℓ_1-norm metric privacy on one-dimensional numerical values, the total variation bound β is:

$$D_1(Laplace(0, 1) \| Laplace(d_{01}, 1)) = 1 - e^{-d_{01}/2}.$$

Table 2: A summary of variation-ratio parameters of ϵ-LDP randomizers. A lower β means a stronger amplification effect.

randomizer	parameter p	parameter β	parameter q
general mechanisms	e^ϵ	$\frac{e^\epsilon-1}{e^\epsilon+1}$	e^ϵ
Laplace mechanism for $[0,1]$ [29]	e^ϵ	$1-e^{-\epsilon/2}$	e^ϵ
PrivUnit mechanism with cap area c [14]	e^ϵ	$\frac{c\cdot(e^\epsilon-1)}{c\cdot e^\epsilon+1-c}$	e^ϵ
general randomized response (GRR) on d options [50]	e^ϵ	$\frac{e^\epsilon-1}{e^\epsilon+d-1}$	e^ϵ
binary RR on d options [26]	e^ϵ	$\frac{e^{\epsilon/2}-1}{e^{\epsilon/2}+1}$	e^ϵ
k-subset on d options [77, 92]	e^ϵ	$\frac{(e^\epsilon-1)\left(\binom{d-1}{k-1}-\binom{d-2}{k-2}\right)}{e^\epsilon\binom{d-1}{k-1}+\binom{d-1}{k}}$	e^ϵ
local hash with length l [83]	e^ϵ	$\frac{e^\epsilon-1}{e^\epsilon+l-1}$	e^ϵ
Hadamard response $(K,s,B=1)$ [1]	e^ϵ	$\frac{s(e^\epsilon-1)/2}{se^\epsilon+K-s}$	e^ϵ
Hadamard response $(K,s,B>1)$ [1]	e^ϵ	$\frac{s(e^\epsilon-1)}{se^\epsilon+K-s}$	e^ϵ
sampling RAPPOR on s in d options [66]	e^ϵ	$\frac{s(e^{\epsilon/2}-1)}{d(e^{\epsilon/2}+1)}$	e^ϵ
Wheel on s in d options with length p [82]	e^ϵ	$\frac{sp(e^\epsilon-1)}{spe^\epsilon+(1-sp)}$	e^ϵ

Table 3: A summary of amplification parameters of $\mathcal{S}(\mathcal{R}(x_1^0),...,\mathcal{R}(x_n))$ and $\mathcal{S}(\mathcal{R}(x_1^1),...,\mathcal{R}(x_n)))$ for local $d_{\mathbb{X}}$-DP randomizers, $d_{01}=d_{\mathbb{X}}(x_1^0,x_1^1)$ and $d_{max}=\max_{x\in\mathbb{X}}\max\{d_{\mathbb{X}}(x_1^0,x),d_{\mathbb{X}}(x_1^1,x)\}$.

randomizer	parameter p	parameter β	parameter q
general mechanisms	$e^{d_{01}}$	$\frac{e^{d_{01}}-1}{e^{d_{01}}+1}$	$e^{d_{max}}$
Laplace mechanism [2], ℓ_1-norm on $\mathbb{R}$	$e^{d_{01}}$	$1-e^{-d_{01}/2}$	$e^{d_{max}}$
planar Laplace [3], ℓ_2-norm on $\mathbb{R}^2$	$e^{d_{01}}$	$2\int_0^{\frac{d_{01}}{2}}\int_{-\infty}^{+\infty}\frac{e^{-\sqrt{(x-\frac{d_{01}}{2})^2+y^2}}}{2\pi}dydx$	$e^{d_{max}}$

Table 4: A summary of amplification parameters of some multi-message shuffle private protocols.

randomizer	parameter p	parameter β	parameter q
Balcer *et al.* [4] with coin prob. p for binary summation	$+\infty$	1	$\max\{\frac{1}{p},\frac{1}{1-p}\}$
Balcer *et al.* [5] with uniform coin for binary summation	$+\infty$	1	2
Cheu *et al.* [21] with flip prob. $f\in[0,0.5]$ on $\{0,1\}^d$	$\frac{(1-f)^2}{f^2}$	$1-2f$	$\frac{1-f}{f}$
Balls-into-bins [61] with d bins (and s special bins)	$+\infty$	1	$\frac{d}{s}$
mixDUMP [58] with flip prob. $f\in[0,\frac{d-1}{d}]$ on d bins	$\frac{(1-f)(d-1)}{f}$	$\frac{(1-f)(d-1)-f}{d-1}$	$(1-f)d$

For the location randomization mechanism with ℓ_2-norm metric privacy over two-dimensional domains, such as the planar Laplace mechanism [3], the probability density $\mathbb{P}[PlanarLaplace(u,1)=x]=\frac{e^{-\|x-u\|_2}}{2\pi}$ for $u,x\in\mathbb{R}^2$, and the total variation bound β is:

$$2\int_0^{d_{01}/2}\int_{-\infty}^{+\infty}(e^{-\sqrt{(x-d_{01}/2)^2+y^2}})/(2\pi)dydx.$$

Multi-message protocols. For multi-message protocols in the literature, we revisit their actual variation-ratio parameters in Table 4. Typically, in these protocols, messages sent by a user can be classified as input-dependent or input-independent, with only one of them being input-dependent. For example, for the user 1 among n' users in [21], $\mathcal{R}_1(x_1)$ is produced by binary randomized response on x_i, while other $m-1$ messages (i.e. $\mathcal{R}_{n'+1}(*),\ldots,\mathcal{R}_{(m-1)\cdot n'+1}(*)$) are independently produced by binary randomized response on a zero vector $\{0\}^d$ with a flip probability of $f\in[0,0.5]$. Similarly, in protocols for binary summation ($x_i\in\{0,1\}$), the input-independent/blanket message can be a Bernoulli variable with biased coin [4] or uniform coin [5]. In balls-into-bins [61], pureDUMP,

and mixDUMP [58], each blanket message is a uniform-random category in $[d]$. The parameters p and β are computed from the input-dependent message $\mathcal{R}_1(x_1)$, while the parameter q is derived from input-independent messages contributed by both user 1 and other users, referred to as blanket messages in [7]. It is important to emphasize that, within these multi-message shuffle protocols, input-independent blanket messages can be viewed as if they were generated by dummy users. Consequently, the number of messages in which an input-dependent message can blend is equivalent to the total count of input-independent messages. For instance, assuming n' users with each user contributing $m-1$ input-independent messages, the term $n-1$ in Theorem 4.7 effectively becomes $n'\cdot(m-1)$. Compared to the privacy guarantees in the original works, our amplification bound for these protocols offers a more than 70% budget savings, as demonstrated by the numerical results in Section 7.2.

Discussion on the generality of our framework for multi-message protocols. We note that our framework applies to all single-message shuffle protocols using local ϵ-DP or metric DP randomizers. While for multi-message protocols in the shuffle model,

based on the correlation between outputted messages, they can be primarily categorized into three main types:

I. The first type of multi-message shuffle protocols is achieved by invoking multiple single-message protocols. This is seen as in the Recursive protocol in [8] and the utility-complexity balanced protocol in [44].

II. The second type involves sending one input-dependent message and multiple independent blanket messages, as seen in [4, 5, 21, 58, 61].

III. The third type involves sending multiple *correlated* messages per user. Protocols falling in this category include [37, 38] and the secret-sharing-style IKOS protocol in [8].

We also envision that there can be multi-message protocols that combines Types I and II, by invoking several protocols each is a multi-message one. Our variation-ratio framework is compatible with protocols categorized under Type I and/or Type II. However, Type III protocols, wherein messages from a single user are correlated, breach the independence assumption of each randomizer $\mathcal{R}_i$ within our framework.

Discussion on variation-ratio parameters. The research on parameters p, β, and q shows patterns that help design protocols for the shuffle model. Single-message shuffle protocols that utilize ϵ_0-LDP randomizers are limited to $q = p = e^{\epsilon_0}$. Therefore, the clone probability $2r = 2\beta p/((p-1)q)$ decreases significantly with ϵ_0, and larger population are required to achieve global (ϵ, δ)-DP. Randomization mechanisms with better utility in the local model often fully exploit privacy constraints and have larger total variation bounds β, resulting in weaker privacy amplification effects (e.g., comparing randomized response with binary RR on 2 options). Multi-message shuffle protocols typically set $q < p$ or even $q \ll p$ to increase the clone probability and reduce the number of messages transmitted per user. For local metric $d_{\mathbb{X}}$-DP randomizers that handle large data domains, d_{max} is often set relatively high, and thus $q \geq p$. When a global privacy target is given, they need to strike a balance between local data utility (which increases with d_{max}) and privacy amplification effects (which decrease with d_{max}).

4.4 Numerical Method for Upper Bounds

This part is dedicated to numerically computing the indistinguishable level between two shuffled message sets $\mathcal{S}(\mathcal{R}_1(x_1^0), \ldots, \mathcal{R}_n(x_n))$ and $\mathcal{S}(\mathcal{R}_1(x_1^1), \ldots, \mathcal{R}_n(x_n))$, within a given privacy failure probability $\delta \in [0, 1]$. According to Theorem 4.7, the indistinguishable level is upper bounded by the one between binomial counts $P_{p,\beta}^q$ and $Q_{p,\beta}^q$. Our aim is then to find the smallest $\epsilon \in [0, \log p]$ that the corresponding Hockey-stick divergence is no more than the privacy failure parameter:

$$\arg \min_{\epsilon \in [0, \log p]} \max \left[D_{e^\epsilon}(P_{p,\beta}^q \| Q_{p,\beta}^q), D_{e^\epsilon}(Q_{p,\beta}^q \| P_{p,\beta}^q) \right] \leq \delta. \quad (4)$$

Solving this problem directly is infeasible. Fortunately, the divergence monotonically non-increases with ϵ, so one can binary search on ϵ. Plugging in the results from Theorems 4.1 and 4.8, we compute the Hockey-stick divergence given the temporary level ϵ as an expectation of cumulative probabilities. The complete implementation for solving problem (4) is provided Algorithm 1.

Algorithm 1: Efficient search of indistinguishable upper bound of $P_{p,\beta}^q$ and $Q_{p,\beta}^q$

Input: parameter $\delta \in [0, 1]$, number of users n, parameters $p > 1, q \geq 1, \beta \in [0, \frac{p-1}{p+1}]$, number of iterations T.

Output: An upper bound of ϵ_c' that
$$\max \left[D_{\epsilon_c'}(P_{p,\beta}^q \| Q_{p,\beta}^q), D_{\epsilon_c'}(Q_{p,\beta}^q \| P_{p,\beta}^q) \right] \leq \delta \text{ holds.}$$

1   $\alpha = \frac{\beta}{p-1}$, $r = \frac{\alpha p}{q}$
2   **Procedure** Delta(ϵ')
3     $\delta' \leftarrow 0$
4     $w_c = \binom{n-1}{c}(2r)^c(1-2r)^{n-1-c}$
5     $low_c = \frac{(e^{\epsilon'}p-1)\alpha c + (e^{\epsilon'}-1)f}{\alpha(e^{\epsilon'}+1)(p-1)}$
6   **for** $c \in [0, n]$ **do**
7       $\delta' \leftarrow \delta' + w_c((p - e^\epsilon)\alpha \cdot \underset{c,1/2}{\text{CDF}}[\lceil low_{c+1} - 1 \rceil, c] + (1 - pe^\epsilon)\alpha \cdot \underset{c,1/2}{\text{CDF}}[\lceil low_{c+1} \rceil, c] + (1 - e^\epsilon)(1 - \alpha - p\alpha) \cdot \underset{c,1/2}{\text{CDF}}[\lceil low_c \rceil, c])$
8   **end**
9   **return** δ'
10   $\epsilon_L \leftarrow 0$, $\epsilon_H \leftarrow \log(p)$
11   **for** $t \in [T]$ **do**
12     $\epsilon_t \leftarrow \frac{\epsilon_L + \epsilon_H}{2}$
13   **if** Delta(ϵ_t) $> \delta$ **then**
14       $\epsilon_L \leftarrow \epsilon_t$
15   **else**
16       $\epsilon_H \leftarrow \epsilon_t$
17   **end**
18   **end**
19   **return** ϵ_H

Due to symmetry, the divergence $D_{e^\epsilon}(P_{p,\beta}^q \| Q_{p,\beta}^q)$ is always equal to $D_{e^\epsilon}(Q_{p,\beta}^q \| P_{p,\beta}^q)$. Thus, Algorithm 1 only needs to compute one of them. We set the binary search upper bound, ϵ_H, to $\log(p)$, given that the differing data item x_1 is at least protected by $\mathcal{R}_1$ with an indistinguishability level of $\log(p)$. The algorithm includes a sub-procedure, Delta(ϵ), computing $D_{e^\epsilon}(P_{p,\beta}^q \| Q_{p,\beta}^q)$ in $\tilde{O}(n)$ time. Hence, the total computational complexity is $\tilde{O}(n \cdot T)$, with T being the number of binary search iterations, influencing the precision of the numerical indistinguishability level.

5 AMPLIFICATION LOWER BOUNDS

In this section, we aim to establish lower bounds for privacy amplification via shuffling. These bounds serve to demonstrate the tightness of the upper bounds derived in the preceding sections. Aiming at maximizing the divergence in observable shuffled messages given two neighboring datasets, our approach involves selecting elements $y \in \mathbb{Y}$ in the message space where the probabilities $\mathbb{P}[\mathcal{R}(x_1^0) = y]$ and $\mathbb{P}[\mathcal{R}(x_1^1) = y]$ differ. We then choose the worst-case data $x^* \in \mathbb{X}$ for other users to maximize the *expected* probability ratio over differed elements, and finally, we observe the number of occurrences of these differing elements in shuffled messages. For succinctness, we summarize the occurrences of these elements and summarize them into two binomial counts, similar to the technique used in Theorem 4.7. The resulting amplification lower bound is presented in Theorem 5.1, indicating the (worst-case) divergence

of shuffled messages is lower bounded by the divergence between two binomial counts $P_{p_0,\beta}^{q_0,q_1}$ and $Q_{p_0,\beta}^{q_0,q_1}$.

Theorem 5.1 (Privacy amplification lower bounds). *Given $x_1^0, x_1^1 \in \mathbb{X}$ and local randomizers $\{\mathcal{R}_1, \mathcal{R}_2\}$ that have finite output domain $\mathbb{Y}$, let p_0 denote $\frac{\sum_{y \in \mathbb{Y}} [\![\mathbb{P}[\mathcal{R}_1(x_1^1)=y] > \mathbb{P}[\mathcal{R}_1(x_1^0)=y]]\!] \cdot \mathbb{P}[\mathcal{R}_1(x_1^1)=y]}{\sum_{y \in \mathbb{Y}} [\![\mathbb{P}[\mathcal{R}_1(x_1^1)=y] > \mathbb{P}[\mathcal{R}_1(x_1^0)=y]]\!] \cdot \mathbb{P}[\mathcal{R}_1(x_1^0)=y]}$, let β denote $D_{e^0}(\mathcal{R}_1(x_1^1) \| \mathcal{R}_1(x_1^0))$. Find an $x^* \in \mathbb{X}$ such that:*

$$x^* = \arg\max_{x \in \mathcal{X}} \min$$

$$\left\{ \frac{\sum_{y \in \mathcal{Y}} \left[\mathbb{P}[\mathcal{R}_1(x_1^1) = y] > \mathbb{P}[\mathcal{R}_1(x_1^0) = y] \right] \cdot \mathbb{P}[\mathcal{R}_1(x_1^1) = y]}{\sum_{y \in \mathcal{Y}} \left[\mathbb{P}[\mathcal{R}_1(x_1^1) = y] > \mathbb{P}[\mathcal{R}_1(x_1^0) = y] \right] \cdot \mathbb{P}[\mathcal{R}_2(x) = y]}, \right.$$

$$\left. \frac{\sum_{y \in \mathcal{Y}} \left[\mathbb{P}[\mathcal{R}_1(x_1^1) = y] < \mathbb{P}[\mathcal{R}_1(x_1^0) = y] \right] \cdot \mathbb{P}[\mathcal{R}_1(x_1^0) = y]}{\sum_{y \in \mathcal{Y}} \left[\mathbb{P}[\mathcal{R}_1(x_1^1) = y] < \mathbb{P}[\mathcal{R}_1(x_1^0) = y] \right] \cdot \mathbb{P}[\mathcal{R}_2(x) = y]} \right\},$$

let q_1 denote $\frac{\sum_{y \in \mathbb{Y}} [\![\mathbb{P}[\mathcal{R}_1(x_1^1)=y] > \mathbb{P}[\mathcal{R}_1(x_1^0)=y]]\!] \cdot \mathbb{P}[\mathcal{R}_1(x_1^1)=y]}{\sum_{y \in \mathbb{Y}} [\![\mathbb{P}[\mathcal{R}_1(x_1^1)=y] > \mathbb{P}[\mathcal{R}_1(x_1^0)=y]]\!] \cdot \mathbb{P}[\mathcal{R}_2(x^*)=y]}$ and q_0 denote $\frac{\sum_{y \in \mathbb{Y}} [\![\mathbb{P}[\mathcal{R}_1(x_1^1)=y] < \mathbb{P}[\mathcal{R}_1(x_1^0)=y]]\!] \cdot \mathbb{P}[\mathcal{R}_1(x_1^0)=y]}{\sum_{y \in \mathbb{Y}} [\![\mathbb{P}[\mathcal{R}_1(x_1^1)=y] < \mathbb{P}[\mathcal{R}_1(x_1^0)=y]]\!] \cdot \mathbb{P}[\mathcal{R}_2(x^*)=y]}$. Let $\alpha = \frac{\beta'}{(p-1)}$, $r_0 = \frac{\alpha p_0}{q_0}$ and $r_1 = \frac{\alpha p_0}{q_1}$, and $C \sim Binom(n-1, r_0 + r_1)$, $A \sim Binom(C, r_0/(r_0+r_1))$, and $\Delta_1 = Bernoulli(p_0\alpha), \Delta_2 = Bernoulli(1-\Delta_1, \alpha/(1-p_0\alpha))$. Let $P_{p_0,\beta}^{q_0,q_1}$ denote $(A+\Delta_1, C-A+\Delta_2), Q_{p_0,\beta}^{q_0,q_1}$ denote $(A+\Delta_2, C-A+\Delta_1)$, then there exists $x_2, ..., x_n \in \mathbb{X}$ such that:

$$D(\mathcal{S}(\mathcal{R}_1(x_1^0), ..., \mathcal{R}_2(x_n)) \| \mathcal{S}(\mathcal{R}_1(x_1^1), ..., \mathcal{R}_2(x_n))) \geq D(P_{p_0,\beta}^{q_0,q_1} \| Q_{p_0,\beta}^{q_0,q_1}).$$

According to the definition of binomial counts $P_{p_0,\beta}^{q_0,q_1}, Q_{p_0,\beta}^{q_0,q_1}$ in the lower bounding Theorem 5.1 and $P_{p,\beta}^q, Q_{p,\beta}^q$ in the upper bounding Theorem 4.7, when $p_0 = p$ and $q_0 = q_1 = q$, the upper bound matches lower bound and is thus exactly tight. This requires that the expected probability ratio p_0 equals to the maximum ratio p. This can be met by randomizers that employ an extremal probability design [51] that probability ratio $\mathbb{P}[\mathcal{R}(x_i) = y]/\mathbb{P}[\mathcal{R}(x_{i'}) = y]$ must belongs to $\{1, e^\epsilon, e^{-\epsilon}\}$ for all $i, i' \in [n]$, $x_i, x_{i'} \in \mathbb{Y}$, and $y \in \mathbb{Y}$. Besides, the matching condition requires that the expected probability ratios q_0 and q_1 equal to the maximum ratio q.

Most state-of-the-art ϵ-LDP randomizers adhere to the mentioned criteria. Notable examples include the generalized randomized response [50] for more than 2 options, the k-subset mechanism [77] with cardinality $k \leq 2$, local hash [83] with length $l \geq 3$, Hardamard response [1], PrivUnit [14] with cap area $c \leq 1/2$, PCKV-GRR [45], and Wheel mechanism [82] with length $p \geq 1/(2s)$. As a result, the amplification upper bounds for these randomizers match the lower bounds and are precisely tight. In a similar vein, the privacy amplification upper bounds for several recent multi-message protocols are also exactly tight. Examples include Cheu *et al.* [21] with $d \geq 3$, Balls-into-bins [61] with $d \geq 3s$, pureDUMP and mixDUMP [58] with $d \geq 3$ (listed in Table 4). Similar to the upper bound, there exists $\tilde{O}(n)$-complexity algorithms to compute the numerical value of the lower bound. For other randomizers that fail to meet the aforementioned criteria, such as the Laplace and k-subset mechanism [77] with cardinality $k \geq 3$, the expected ratios p_0 and q_0, q_1 are close to the maximum ratio p and q respectively. Consequently, the upper bound remains roughly tight.

6 PARALLEL COMPOSITION IN THE SHUFFLE MODEL

Most data analysis tasks involve multiple estimation targets. In the local setting of DP, it is common practice to partition the entire user population into multiple non-overlapping subsets and allow each subset to handle one estimation query with full budget ϵ_0. This technique is used to achieve better utility compared to dividing the privacy budget ϵ_0. Examples of tasks that utilize this approach include heavy hitter estimation [10, 66, 86], multi-dimensional data publication [69, 75, 84], frequent itemset mining [85], range queries [24], marginal queries [22], data synthesis [87, 93], and machine learning [34, 48, 89]. This technique aligns with the parallel composition theorem of DP in the centralized setting [27].

In the shuffle model of DP, a naive approach to handling multiple queries is to divide the population into K cohorts and separately amplify privacy with about n/K users for each query. A more effective approach is to allow each user to randomly select one query from all K queries using a common distribution $P_k \in \Delta_K$, and to contribute to the selected query in parallel. We illustrate this parallel approach in Algorithm 2. Since all base mechanisms $\mathcal{M}_k$ (for $k \in [K]$) satisfy ϵ_0-LDP, the overall algorithm satisfies ϵ_0-LDP.

Algorithm 2: Parallel local randomizer

Params: Number of queries K, a probability distribution $P_k : [K] \mapsto [0, 1]$, local base randomizers $\{\mathcal{M}_k : \mathbb{X} \mapsto \mathbb{Y}_k\}_{k \in [K]}$ each satisfies ϵ_0-LDP and corresponds to one query.

Input: An input $x \in \mathbb{X}$.

Output: An output z satisfies ϵ_0-LDP.

1 sample $k \sim P_k$
2 $y \leftarrow \mathcal{M}_k(x)$
3 **return** y

In light of the fact that each user draws a sample k from the identical query distribution P_k, all users adopt the same randomization algorithm (Algorithm 2) that satisfies ϵ_0-LDP. As a result, the privacy amplification via shuffling discussed in the previous sections remains valid with n users amplify privacy together. We refer to Algorithm 2 as $\mathcal{R}$. Drawing on the $(e^{\epsilon_0}, \frac{e^{\epsilon_0}-1}{e^{\epsilon_0}+1})$-variation property and the e^{ϵ_0}-ratio property of arbitrary identical ϵ_0-LDP randomizers, we can straightforwardly conclude that:

$$D(\mathcal{S}(\mathcal{R}(x_1^0), ..., \mathcal{R}(x_n)) \| \mathcal{S}(\mathcal{R}(x_1^1), ..., \mathcal{R}(x_n)))$$

$$\leq D\left(P_{e^{\epsilon_0}, \frac{e^{\epsilon_0}-1}{e^{\epsilon_0}+1}}^{e^{\epsilon_0}} \| Q_{e^{\epsilon_0}, \frac{e^{\epsilon_0}-1}{e^{\epsilon_0}+1}}^{e^{\epsilon_0}} \right),$$

which we refer to as the basic parallel composition theorem.

By exploiting the connection between total variation bounds and indistinguishable levels established in this study, we can derive a stronger privacy guarantee for Algorithm 2. Specifically, the total variation bound of the parallel local randomizer is bounded by the expectation of the base randomizers' total variation bound. We present this improved result in Theorem 6.1 and refer to it as the advanced parallel composition in the shuffle model.

Theorem 6.1 (Advanced parallel composition in the shuffle model). *Assume $\mathcal{M}_k$ satisfy the $(e^{\epsilon_0}, \beta_k)$-variation property, and let*

$\mathcal{R}$ denote the Algorithm 2, then for any inputs $x_1^0, x_1^1, x_2, ..., x_n \in \mathbb{X}$:

$$D(\mathcal{S}(\mathcal{R}(x_1^0), ..., \mathcal{R}(x_n)) \| \mathcal{S}(\mathcal{R}(x_1^1), ..., \mathcal{R}(x_n))) \leq D\left(P_{e^{\epsilon_0}, \bar{\beta}}^{e^{\epsilon_0}} \| Q_{e^{\epsilon_0}, \bar{\beta}}^{e^{\epsilon_0}}\right)$$

where $\bar{\beta} = \sum_{k' \in [K]} \beta_{k'} \cdot \mathbb{P}[P_k = k']$.

PROOF. To establish the theorem, it is sufficient to demonstrate that $\mathcal{R}$ satisfies the $(e^{\epsilon_0}, \bar{\beta})$-variation property. The fact that $\mathcal{R}$ satisfies $D_{e^{\epsilon_0}}(\mathcal{R}(x_1^0) \| \mathcal{R}(x_1^1)) = 0$ is a direct consequence of the ϵ_0-LDP guarantee of $\mathcal{R}$. Moreover, applying the definition of the total variation (or Hockey-stick divergence), we obtain:

$$D_1(\mathcal{R}(x_1^0) \| \mathcal{R}(x_1^1)) \leq \sum_{k' \in [K]} \mathbb{P}[P_k = k'] \cdot D_1(\mathcal{M}_{k'}(x_1^0) \| \mathcal{M}_{k'}(x_1^1))$$
$$\leq \sum_{k' \in [K]} \mathbb{P}[P_k = k'] \cdot \beta_{k'}.$$

$\square$

7 NUMERICAL RESULTS

In this section, we present evaluations of the proposed variation-ratio framework for both single-message and multi-message protocols in the shuffle model. Our main focus is to demonstrate the effectiveness and efficiency of numerical upper bounds. We also validate the effectiveness of the closed-form bounds presented in Theorems 4.2 and 4.3. Additionally, we demonstrate the performance improvements achieved through the advanced parallel composition presented in Theorem 6.1, taking private range queries in the shuffle model as a concrete example.

7.1 On Single-message Protocols

To evaluate the effectiveness of the proposed variation-ratio framework for privacy amplification on LDP randomizers, we compare it with existing amplification bounds such as the privacy blanket [7], clone reduction [32], and the stronger clone reduction [33]. We specifically consider two state-of-the-art LDP randomizers for discrete distribution estimation: the subset selection mechanism [77, 92] and the optimal local hash [83]. For the privacy blanket method, we present the tighter bound between the upper bounds given by "Hoeffding, Generic" and "Bennett, Generic" for general LDP randomizers, denoted as *privacy-blanket, general*. We also present the tighter bound between the upper bounds given by "Hoeffding" and "Bennett" with randomizer-specific parameters, i.e., with total variation similarity $\gamma = \binom{d}{k} / (e^{\epsilon_0} \binom{d-1}{k-1} + \binom{d-1}{k})$ for the subset mechanism and total variation similarity $\gamma = \frac{l}{e^{\epsilon_0} + l - 1}$ for the optimal local hash, denoted as *privacy-blanket, specific*. All presented results are numerical amplification upper bounds (except the classical EFMRTT19 [30] providing only closed-form bounds).

We use *amplification ratio* to measure the effectiveness of privacy amplification. It is defined as the ratio of the local budget ϵ_0 to the amplified privacy budget ϵ:

$$\text{amplification ratio} = \epsilon_0 / \epsilon.$$

We present the amplification ratio results for the subset and optimal local hash mechanisms in Figures 1 and 2, respectively. The results demonstrate that our variation-ratio analyses can save about 30% of the privacy budget for both mechanisms when compared with the best existing bounds. Notably, in the optimal local hash mechanism, when $l = e^{\epsilon_0} + 1$ is greater than 2, the amplification upper bounds obtained from our variation-ratio framework are tight, matching the lower bounds presented in Section 5.

7.2 On Multi-message Protocols

To assess the efficacy of variation-ratio analyses for multi-message protocols, we apply it to two state-of-the-art histogram estimation protocols, namely Cheu *et al.* [21] and balls-into-bins [61]. We compare the amplified privacy provided by the original works, denoted as ϵ', with that given by variation-ratio analyses, denoted as ϵ. To measure the additional privacy amplification offered by our analyses, we use the *extra amplification ratio*, defined as:

$$\text{extra amplification ratio} = \epsilon'/\epsilon.$$

As depicted in Figures 3 and 4, our numerical bounds, denoted as *variation-ratio*, are significantly tighter and result in a privacy budget savings of approximately 75%. Moreover, we present closed-form bounds from Theorem 4.2 and Theorem 4.3, denoted as *variation-ratio, analytic* and *variation-ratio, asymptotic*, respectively. Both bounds are tighter than those obtained from the original works, and the closed-form bounds from Theorem 4.2 yield more than 50% budget savings.

7.3 On Advanced Parallel Composition

This part evaluates the privacy amplification effects provided by the advanced parallel composition in Theorem 6.1. To illustrate the effectiveness of this technique, we consider range queries over a categorical domain $[1 : d]$, a well-studied problem in the literature [24, 30]. To avoid $\Theta(d)$ errors in estimators, a common practice is to represent categories in a hierarchical form and let each user report one hierarchy level. For categorical domains of size $d = 2^H$, the k-th value in the h-th hierarchy is given by:

$$V_{h,k} = [\![j \in [(k-1) \cdot 2^h : k \cdot 2^h]]\!],$$

where $h \in [0 : H-1]$, $k \in [1 : d/2^h]$, and $[\![ ]\!]$ is the Iverson bracket.

Follow the approach suggested in [24], we assume each user uniformly selects one hierarchy level $h \in [0 : H-1]$ and uses the generalized randomized response mechanism with full budget ϵ_0 to report the one-hot vector $V_{h,*}$ (the generalized randomized response is optimal in the low local privacy regime [92]). Using the basic parallel composition theorem, this parallel local randomizer with H queries implies the following variation-ratio parameters: $p = e^{\epsilon_0}$, $\beta = \frac{e^{\epsilon_0} - 1}{e^{\epsilon_0} + 1}$, and $q = e^{\epsilon_0}$. In contrast, according to Table 2 and the advanced parallel composition in Theorem 6.1, the tighter variation-ratio parameters can be computed as follows: $p = e^{\epsilon_0}$, $q = e^{\epsilon_0}$, and $\beta = \sum_{h \in [0:H-1]} \frac{1}{H} \frac{e^{\epsilon_0} - 1}{e^{\epsilon_0} + d/2^h - 1}$.

We present a comparative analysis of the numerical privacy amplification achieved through advanced parallel composition and basic parallel composition in Figure 5, for the following settings: $d = 64$ or 2048, and $n = 10^4$ or 10^5. Our findings demonstrate that the use of advanced parallel composition results in a reduction of the privacy budget by approximately 75%. Moreover, we investigate the privacy amplification effects of the separate approach wherein non-overlapping users report each hierarchy level separately. We present the amplification results for n/H users using the best-possible parameters: $p = e^{\epsilon_0}$, $q = e^{\epsilon_0}$, and $\beta = \frac{e^{\epsilon_0} - 1}{e^{\epsilon_0} + d - 1}$ (denoted as *separate, best* in Figure 5) or the worst-possible parameters: $p = e^{\epsilon_0}$, $q = e^{\epsilon_0}$, and $\beta = \frac{e^{\epsilon_0} - 1}{e^{\epsilon_0} + 1}$ (denoted as *separate, worst*). Our results show that the advanced parallel composition saves budget by 80%-95% when compared to the separated approach.
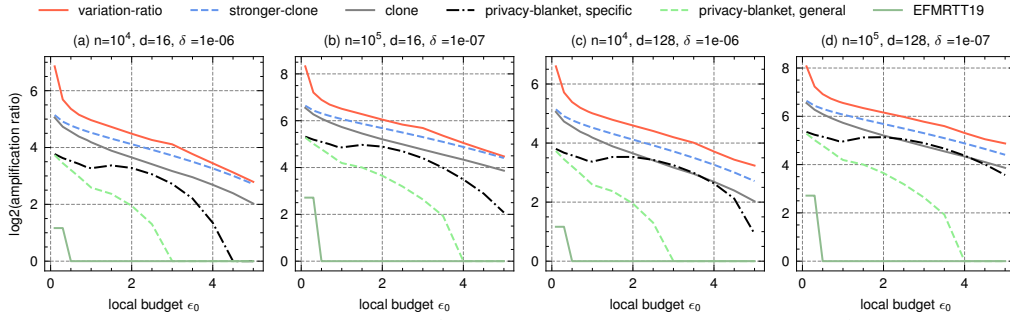
Figure 1: Numerical comparison of amplification effects (base 2 logarithm of amplification ratio) of subset selection mechanism with $n = 10^4$ or 10^5, domain size $d = 16$ or 128, and varying local budget $\epsilon_0 \in [0.1, 5.0]$.

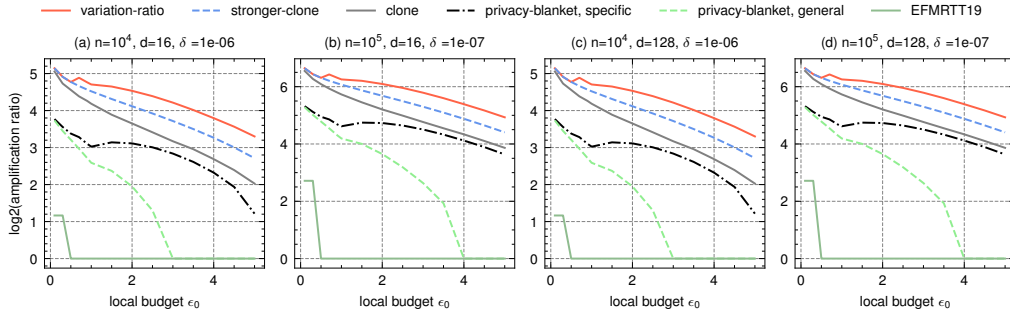**Figure 2: Numerical comparison of amplification effects (base 2 logarithm of amplification ratio) of optimal local hash mechanism with $n = 10^4$ or 10^5, domain size $d = 16$ or 128, and local budget $\epsilon_0 \in [0.1, 5.0]$.**
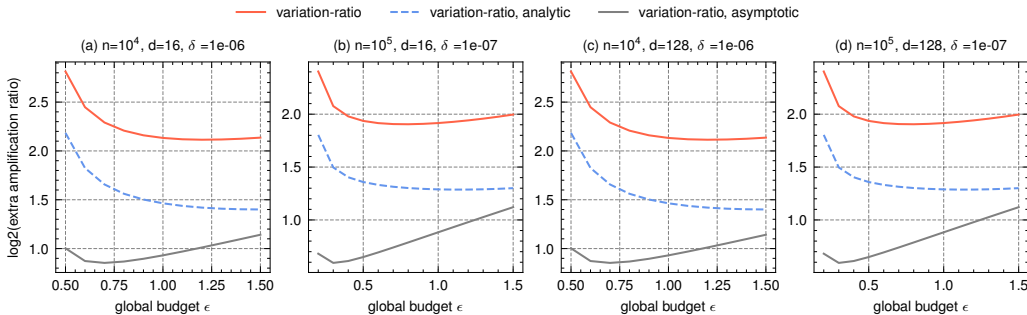

Figure 3: Numerical comparison of amplification effects (base 2 logarithm of extra amplification ratio) of the Cheu _et al._ [21] multi-message protocol with $n = 10^4$ or 10^5, domain size $d = 16$ or 128, and varying global budget $\epsilon' \in [0.01, 1.5]$.

7.4 Efficiency Evaluation

We implement Theorem 4.1 and Algorithm 1 in Python 3.8 and execute it on a desktop computer with Intel Core i7-10700KF @3.8GHz and 32GB memory. In Table 5, we present the results for $T = 20$ or $T = 10$ (T is the number of binary search iterations on amplified level ϵ, see Section 4.4), where we vary the local budget ϵ_0 and number of users n. The numerical privacy amplification bounds presented in the table correspond to general LDP mechanisms (refer to Table 2 for variation-ratio parameters). Our results demonstrate that the running time of the implementation is less sensitive to the local budget ϵ_0, and mainly depends on the population size n and number of iterations T. Furthermore, the running time grows

linearly with n, and we can obtain tight numerical privacy amplification bounds within a dozen seconds even when n is extremely large (e.g., $n = 10^8$). Our results also show that one can trade tightness for computational efficiency by choosing a smaller value for T. Specifically, comparing the results obtained for $T = 10$ with those for $T = 20$, we observe that the latter provides slightly tighter bounds, but at the cost of much longer running time.

8 CONCLUSION

This work has presented a unified, tight, easy-to-use, and efficient framework for analyzing privacy amplification within the emerging shuffle model. The framework has been shown to provide tight
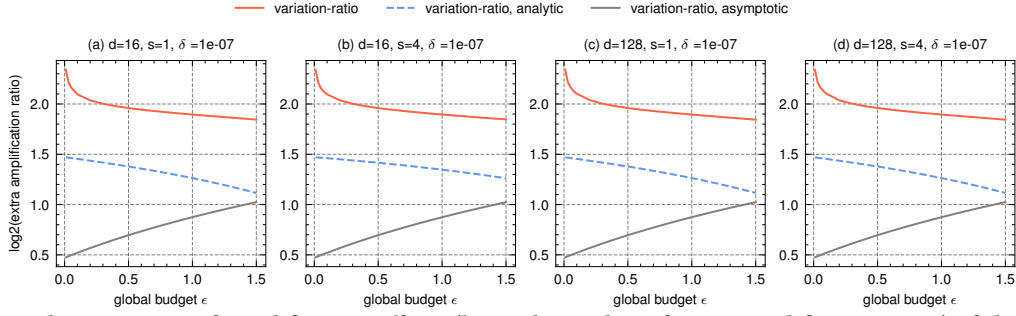
Figure 4: Numerical comparison of amplification effects (base 2 logarithm of extra amplification ratio) of the balls-into-bins multi-message protocol with $n = \frac{32 \log(2/\delta)d}{\epsilon'^2 s}$ [61] and varying global budget $\epsilon' \in [0.01, 1.5]$.
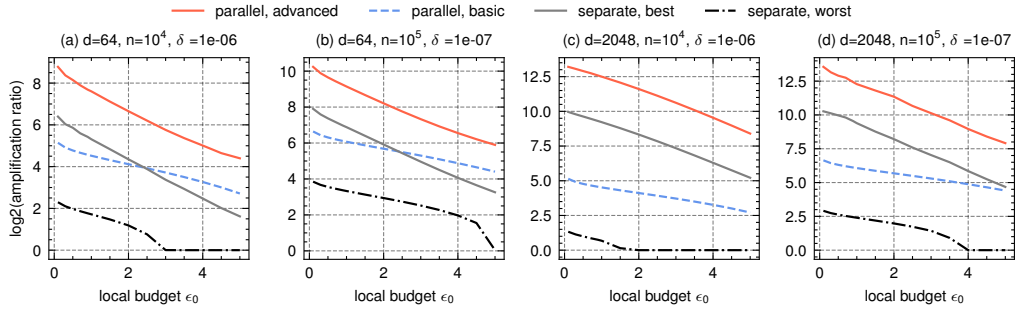

Figure 5: Numerical comparison of amplification effects (base 2 logarithm of amplification ratio) of separated approach, basic parallel composition, and advanced parallel composition.

Table 5: Amplification bounds (w.r.t. $\delta = 0.01/n$) and running time comparison for general ϵ_0-LDP randomizers with iterations $T = 20$ or $T = 10$ and varying number of users.

		$T = 20$		$T = 10$	
	number of users n	amplified privacy ϵ	time (seconds)	amplified privacy ϵ	time (seconds)
$\epsilon_0 = 1.0$	10^4	0.0433	0.45	0.0440	2.3
	10^6	0.00503	16.7	0.00586	4.7
	10^8	0.000566	106.2	0.000977	0.43
$\epsilon_0 = 3.0$	10^4	0.227	3.8	0.229	1.5
	10^6	0.0255	13.9	0.0264	2.8
	10^8	0.00283	**117.5**	0.00293	5.2
$\epsilon_0 = 5.0$	10^4	0.743	2.4	0.743	0.9
	10^6	0.0778	9.3	0.0782	3.1
	10^8	0.00853	44.6	0.00977	**7.9**
$\epsilon_0 = 7.0$	10^4	6.99	0.24	6.99	0.31
	10^6	0.224	2.8	0.225	1.2
	10^8	0.0242	17.9	0.0273	5.1

bounds for a wide range of state-of-the-art single-message/multi-message protocols, by translating two novel yet intuitive parameterizations (namely, total variation and probability ratio) about local randomizers into differential privacy levels. Additionally, the framework has induced an advanced parallel composition theorem in the shuffle model, widely applicable to count queries, data synthesis, data mining, and machine learning. Our framework also leads to a fast $\tilde{O}(n)$-complexity procedure for numerical privacy amplification analyses. Comprehensive experiments affirm the effectiveness and efficiency of the proposed framework.

REFERENCES

[1] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. 2018. Hadamard Response: Estimating Distributions Privately, Efficiently, and with Little Communication. *arXiv preprint arXiv:1802.04705* (2018).

[2] Mário Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. 2018. Local differential privacy on metric spaces: optimizing the trade-off with utility. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 262–267.

[3] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: Differential privacy for location-based systems. In *CCS. ACM*.

[4] Victor Balcer and Albert Cheu. 2020. Separating Local & Shuffled Differential Privacy via Histograms. In *1st Conference on Information-Theoretic Cryptography*.

[5] Victor Balcer, Albert Cheu, Matthew Joseph, and Jieming Mao. 2021. Connecting robust shuffle privacy and pan-privacy. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2384–2403.

[6] Borja Balle, Gilles Barthe, and Marco Gaboardi. 2018. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in Neural Information Processing Systems* 31 (2018).

[7] Borja Balle, James Bell, Adria Gascón, and Kobbi Nissim. 2019. The privacy blanket of the shuffle model. *CRYPTO* (2019).

[8] Borja Balle, James Bell, Adria Gascon, and Kobbi Nissim. 2020. Private summation in the multi-message shuffle model. *CCS* (2020).

[9] Ergute Bao, Yin Yang, Xiaokui Xiao, and Bolin Ding. 2021. CGM: an enhanced mechanism for streaming data collection with local differential privacy. *Proceedings of the VLDB Endowment* 14, 11 (2021), 2258–2270.

[10] Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Guha Thakurta. 2017. Practical locally private heavy hitters. *Advances in Neural Information Processing Systems* 30 (2017).

[11] James Bennett, Stan Lanning, et al. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*, Vol. 2007. 35.

[12] Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. 2018. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984* (2018).

[13] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. 2017. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th symposium on operating systems principles*. 441–459.

[14] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 1175–1191.

[15] Luca Bonomi and Li Xiong. 2013. Mining frequent patterns with differential privacy. *Proceedings of the VLDB Endowment* 6, 12 (2013), 1422–1427.

[16] Igor Calzada. 2022. Citizens' data privacy in China: The state of the art of the Personal Information Protection Law (PIPL). *Smart Cities* 5, 3 (2022), 1129–1150.

[17] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of Differential Privacy using metrics. In *PETS. Springer*.

[18] Rui Chen, Benjamin CM Fung, Philip S Yu, and Bipin C Desai. 2014. Correlated network data publication via differential privacy. *The VLDB Journal* 23 (2014), 653–676.

[19] Rui Chen, Noman Mohammed, Benjamin CM Fung, Bipin C Desai, and Li Xiong. 2011. Publishing set-valued data via differential privacy. *Proceedings of the VLDB Endowment* 4, 11 (2011), 1087–1098.

[20] Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. 2019. Distributed differential privacy via shuffling. *EUROCRYPT* (2019).

[21] Albert Cheu and Maxim Zhilyaev. 2022. Differentially private histograms in the shuffle model from fake users. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 440–457.

[22] Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. 2018. Marginal release under local differential privacy. In *Proceedings of the 2018 International Conference on Management of Data*. ACM, 131–146.

[23] Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. 2019. Answering range queries under local differential privacy. *VLDB* (2019).

[24] Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. 2019. Answering range queries under local differential privacy. *Proceedings of the VLDB Endowment* 12, 10 (2019), 1126–1138.

[25] Linkang Du, Zhikun Zhang, Shaojie Bai, Changchang Liu, Shouling Ji, Peng Cheng, and Jiming Chen. 2021. AHEAD: adaptive hierarchical decomposition for range query under local differential privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 1266–1288.

[26] John C Duchi, Michael I Jordan, and Martin J Wainwright. 2013. Local privacy and statistical minimax rates. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, 429–438.

[27] Cynthia Dwork. 2006. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14,*

[28] *2006, Proceedings, Part II 33*. Springer, 1–12.

[29] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings 5*. Springer, 1–19.

[29] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 265–284.

[30] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. 2019. Amplification by shuffling: From local to central differential privacy via anonymity. *SODA* (2019).

[31] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*. ACM, 1054–1067.

[32] Vitaly Feldman, Audra McMillan, and Kunal Talwar. 2022. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 954–964.

[33] Vitaly Feldman, Audra McMillan, and Kunal Talwar. 2023. Stronger privacy amplification by shuffling for Rényi and approximate differential privacy. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 4966–4981.

[34] Sam Fletcher and Md Zahidul Islam. 2019. Decision tree classification with differential privacy: A survey. *ACM Computing Surveys (CSUR)* 52, 4 (2019), 1–33.

[35] Evrard Garcelon, Kamalika Chaudhuri, Vianney Perchet, and Matteo Pirotta. 2022. Privacy amplification via shuffling for linear contextual bandits. In *International Conference on Algorithmic Learning Theory*. PMLR, 381–407.

[36] Badih Ghazi, Noah Golowich, Ravi Kumar, Rasmus Pagh, and Ameya Velingker. 2021. On the power of multiple anonymous messages: Frequency estimation and selection in the shuffle model of differential privacy. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 463–488.

[37] Badih Ghazi, Ravi Kumar, Pasin Manurangsi, and Rasmus Pagh. 2020. Private counting from anonymous messages: Near-optimal accuracy with vanishing communication overhead. In *International Conference on Machine Learning*. PMLR, 3505–3514.

[38] Badih Ghazi, Ravi Kumar, Pasin Manurangsi, Rasmus Pagh, and Amer Sinha. 2021. Differentially private aggregation in the shuffle model: Almost central accuracy in almost a single message. In *International Conference on Machine Learning*. PMLR, 3692–3701.

[39] Badih Ghazi, Pasin Manurangsi, Rasmus Pagh, and Ameya Velingker. 2020. Private aggregation from fewer anonymous messages. In *Advances in Cryptology–EUROCRYPT 2020: 39th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zagreb, Croatia, May 10–14, 2020, Proceedings, Part II 30*. Springer, 798–827.

[40] Antonious Girgis, Deepesh Data, and Suhas Diggavi. 2021. Renyi differential privacy of the subsampled shuffle model in distributed learning. *Advances in Neural Information Processing Systems* 34 (2021), 29181–29192.

[41] Antonious Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. 2021. Shuffled Model of Differential Privacy in Federated Learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2521–2529.

[42] Antonious M Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. 2021. Shuffled model of federated learning: Privacy, accuracy and communication trade-offs. *IEEE journal on selected areas in information theory* 2, 1 (2021), 464–478.

[43] Antonious M Girgis, Deepesh Data, Suhas Diggavi, Ananda Theertha Suresh, and Peter Kairouz. 2021. On the rényi differential privacy of the shuffle model. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2321–2341.

[44] Antonious M Girgis and Suhas Diggavi. 2023. Distributed Mean Estimation for Multi-Message Shuffled Privacy. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*.

[45] Xiaolan Gu, Ming Li, Yueqiang Cheng, Li Xiong, and Yang Cao. 2020. PCKV: Locally Differentially Private Correlated Key-Value Data Collection with Optimized Utility. *USENIX Security* (2020).

[46] Charles Hirschman, Richard Alba, and Reynolds Farley. 2000. The meaning and measurement of race in the US census: Glimpses into the future. *Demography* 37, 3 (2000), 381–393.

[47] Xueyang Hu, Mingxuan Yuan, Jianguo Yao, Yu Deng, Lei Chen, Qiang Yang, Haibing Guan, and Jia Zeng. 2015. Differential privacy in telco big data platform. *Proceedings of the VLDB Endowment* 8, 12 (2015), 1692–1703.

[48] Geetha Jagannathan, Krishnan Pillaipakkamnatt, and Rebecca N Wright. 2009. A practical differentially private random decision tree classifier. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 114–121.

[49] Noah Johnson, Joseph P Near, and Dawn Song. 2018. Towards practical differential privacy for SQL queries. *Proceedings of the VLDB Endowment* 11, 5 (2018), 526–539.

[50] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. 2016. Discrete Distribution Estimation under Local Privacy. In *International Conference on Machine Learning*. 2436–2444.

[51] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2014. Extremal mechanisms for local differential privacy. *Advances in neural information processing systems* 27 (2014).

[52] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2011. What can we learn privately? *SIAM J. Comput.* 40, 3 (2011), 793–826.

[53] Antti Koskela, Mikko A Heikkilä, and Antti Honkela. 2022. Numerical Accounting in the Shuffle Model of Differential Privacy. *Transactions on Machine Learning Research* (2022).

[54] Antti Koskela, Joonas Jälkö, and Antti Honkela. 2020. Computing tight differential privacy guarantees using fft. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2560–2569.

[55] Ios Kotsogiannis, Yuchao Tao, Xi He, Maryam Fanaeepour, Ashwin Machanavajjhala, Michael Hay, and Gerome Miklau. 2019. Privatesql: a differentially private sql query engine. *Proceedings of the VLDB Endowment* 12, 11 (2019), 1371–1384.

[56] Chao Li, Michael Hay, Gerome Miklau, and Yue Wang. 2014. A Data-and Workload-Aware Algorithm for Range Queries Under Differential Privacy. *Proceedings of the VLDB Endowment* 7, 5 (2014).

[57] Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. 2015. The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB journal* 24 (2015), 757–781.

[58] Xiaochen Li, Weiran Liu, Hanwen Feng, Kunzhe Huang, Yuke Hu, Jinfei Liu, Kui Ren, and Zhan Qin. 2023. Privacy enhancement via dummy points in the shuffle model. *IEEE Transactions on Dependable and Secure Computing* (2023).

[59] Junxu Liu, Jian Lou, Li Xiong, Jinfei Liu, and Xiaofeng Meng. 2021. Projected federated averaging with heterogeneous differential privacy. *Proceedings of the VLDB Endowment* 15, 4 (2021), 828–840.

[60] Vincent Liu, Mark A Musen, and Timothy Chou. 2015. Data breaches of protected health information in the United States. *JAMA* 313, 14 (2015), 1471–1473.

[61] Qiyao Luo, Yilei Wang, and Ke Yi. 2022. Frequency Estimation in the Shuffle Model with Almost a Single Message. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2219–2232.

[62] Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*. IEEE, 263–275.

[63] Ilya Mironov, Omkant Pandey, Omer Reingold, and Salil Vadhan. 2009. Computational differential privacy. In *Advances in Cryptology-CRYPTO 2009: 29th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 16-20, 2009. Proceedings*. Springer, 126–142.

[64] Thông T Nguyên, Xiaokui Xiao, Yin Yang, Siu Cheung Hui, Hyejin Shin, and Junbum Shin. 2016. Collecting and analyzing data from smart device users with local differential privacy. *arXiv preprint arXiv:1606.05053* (2016).

[65] RB Paris. 2010. Incomplete beta functions. *NIST Handbook of Mathematical Functions,, ed. Frank WJ Olver, Daniel M. Lozier, Ronald F. Boisvert and Charles W. Clark. Cambridge University Press. http://dlmf. nist. gov/8.17* (2010).

[66] Zhan Qin, Yin Yang, Ting Yu, Issa Khalil, Xiaokui Xiao, and Kui Ren. 2016. Heavy hitter estimation over set-valued data with local differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 192–203.

[67] Wullianallur Raghupathi and Viju Raghupathi. 2014. Big data analytics in healthcare: Promise and potential. *Health information science and systems* 2 (2014), 1–10.

[68] Xuebin Ren, Liang Shi, Weiren Yu, Shusen Yang, Cong Zhao, and Zongben Xu. 2022. LDP-IDS: Local differential privacy for infinite data streams. In *Proceedings of the 2022 international conference on management of data*. 1064–1077.

[69] Xuebin Ren, Chia-Mu Yu, Weiren Yu, Shusen Yang, Xinyu Yang, Julie A McCann, and S Yu Philip. 2018. LoPub: high-dimensional crowdsourced data publication with local differential privacy. *IEEE Transactions on Information Forensics and Security* 13, 9 (2018), 2151–2166.

[70] Jay Tenenbaum, Haim Kaplan, Yishay Mansour, and Uri Stemmer. 2021. Differentially private multi-armed bandits in the shuffle model. *Advances in Neural Information Processing Systems* 34 (2021), 24956–24967.

[71] Stacey Truex, Ling Liu, Ka-Ho Chow, Mehmet Emre Gursoy, and Wenqi Wei. 2020. LDP-Fed: Federated learning with local differential privacy. In *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*. 61–66.

[72] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.

[73] Di Wang, Marco Gaboardi, and Jinhui Xu. 2018. Empirical risk minimization in non-interactive local differential privacy revisited. *Advances in Neural Information Processing Systems* 31 (2018).

[74] Di Wang and Jinhui Xu. 2019. On sparse linear regression in the local differential privacy model. In *International Conference on Machine Learning*. PMLR, 6628–6637.

[75] Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. 2019. Collecting and analyzing multidimensional data with local differential privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 638–649.

[76] Shaowei Wang, Liusheng Huang, Yiwen Nie, Pengzhan Wang, Hongli Xu, and Wei Yang. 2018. PrivSet: Set-Valued Data Analyses with Locale Differential Privacy. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 1088–1096.

[77] Shaowei Wang, Liusheng Huang, Yiwen Nie, Xinyuan Zhang, Pengzhan Wang, Hongli Xu, and Wei Yang. 2019. Local differential private data aggregation for discrete distribution estimation. *IEEE Transactions on Parallel and Distributed Systems* 30, 9 (2019), 2046–2059.

[78] Shaowei Wang, Jin Li, Yuntong Li, Xuandi Luo, Wei Yang, Hongyang Yan, and Changyu Dong. 2023. Metric differential privacy in the shuffle model. *(to be released)* (2023).

[79] Shaowei Wang, Jin Li, Yuqiu Qian, Jiachun Du, Wenqing Lin, and Wei Yang. 2021. Hiding Numerical Vectors in Local Private and Shuffled Messages.. In *IJCAI*. 3706–3712.

[80] Shaowei Wang, Xuandi Luo, Yuqiu Qian, Jiachun Du, Wenqing Lin, and Wei Yang. 2022. Analyzing Preference Data With Local Privacy: Optimal Utility and Enhanced Robustness. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[81] Shaowei Wang, Xuandi Luo, Yuqiu Qian, Youwen Zhu, Kongyang Chen, Qi Chen, Bangzhou Xin, and Wei Yang. 2023. Shuffle Differential Private Data Aggregation for Random Population. *IEEE Transactions on Parallel and Distributed Systems* (2023).

[82] Shaowei Wang, Yuqiu Qian, Jiachun Du, Wei Yang, Liusheng Huang, and Hongli Xu. 2020. Set-valued data publication with local privacy: tight error bounds and efficient mechanisms. *Proceedings of the VLDB Endowment* 13, 8 (2020), 1234–1247.

[83] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. 2017. Locally differentially private protocols for frequency estimation. In *Proc. of the 26th USENIX Security Symposium*. 729–745.

[84] Tianhao Wang, Bolin Ding, Jingren Zhou, Cheng Hong, Zhicong Huang, Ninghui Li, and Somesh Jha. 2019. Answering multi-dimensional analytical queries under local differential privacy. In *Proceedings of the 2019 International Conference on Management of Data*. 159–176.

[85] Tianhao Wang, Ninghui Li, and Somesh Jha. 2018. Locally differentially private frequent itemset mining. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 127–143.

[86] Tianhao Wang, Ninghui Li, and Somesh Jha. 2019. Locally differentially private heavy hitter identification. *IEEE Transactions on Dependable and Secure Computing* 18, 2 (2019), 982–993.

[87] Teng Wang, Xinyu Yang, Xuebin Ren, Wei Yu, and Shusen Yang. 2019. Locally private high-dimensional crowdsourced data release based on copula functions. *IEEE Transactions on Services Computing* 15, 2 (2019), 778–792.

[88] Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* 60, 309 (1965), 63–69.

[89] Bangzhou Xin, Wei Yang, Shaowei Wang, and Liusheng Huang. 2019. Differentially private greedy decision forest. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2672–2676.

[90] Min Xu, Bolin Ding, Tianhao Wang, and Jingren Zhou. 2020. Collecting and analyzing data jointly from multiple services under local differential privacy. *Proceedings of the VLDB Endowment* 13, 12 (2020), 2760–2772.

[91] Min Xu, Tianhao Wang, Bolin Ding, Jingren Zhou, Cheng Hong, and Zhicong Huang. 2019. Dpsaas: Multi-dimensional data sharing and analytics as services under local differential privacy. *Proceedings of the VLDB Endowment* 12, 12 (2019), 1862–1865.

[92] Min Ye and Alexander Barg. 2018. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory* (2018).

[93] Zhikun Zhang, Tianhao Wang, Ninghui Li, Shibo He, and Jiming Chen. 2018. Calm: Consistent adaptive local marginal for marginal release under local differential privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 212–229.

[94] Ying Zhao and Jinjun Chen. 2022. A survey on differential privacy for unstructured data content. *ACM Computing Surveys (CSUR)* 54, 10s (2022), 1–28.

[95] Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. 2022. Optimal accounting of differential privacy via characteristic function. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4782–4817.