# Shuffle Differential Private Data Aggregation for Random Population

Shaowei Wang, Xuandi Luo, Yuqiu Qian, Youwen Zhu, Kongyang Chen, Qi Chen, Bangzhou Xin, and Wei Yang, *Member, IEEE*

***Abstract*—Bridging the advantages of differential privacy in both centralized model (i.e., high accuracy) and local model (i.e., minimum trust), the shuffle privacy model has potential applications in many privacy-sensitive scenarios, such as mobile user data aggregation and federated learning. Since messages from users are anonymized by semi-trusted shufflers (e.g., anonymous channels, edge servers), every user could hide message among other users' messages and inject only part of noises (a.k.a. privacy amplification). However, existing works assume that the participating user population is known in advance, which is unrealistic for dynamic environments (e.g., mobile computing, vehicular networks). In this work, we study the shuffle privacy model with a random participating population, and give privacy amplification bounds for population size with commonly encountered binomial, Poisson, sub-Gaussian distribution and etc. For further improving accuracy, we formulate and derive optimal dummy sizes for both non-adaptive and adaptive dummies. Finally, to break the error barrier due to the constraint of sending one single message per user, we design a multi-message shuffle private protocol supporting random population. Experiment results show that our approaches reduce more than 60% error when compared to the local model and naive approaches. We hope this work provides tailored solutions of shuffle privacy for dynamic mobile/distributed computing.**

***Index Terms*—Data aggregation, data privacy, differential privacy, shuffle privacy, statistical estimation.**

## I. Introduction

Since its advent for privacy-preserving statistical queries in databases, differential privacy [1] has far been applied to numerous areas, such as genetic/medical data analyses [2] and federated machine learning [3]. Specifically, the local model [4] of differential privacy allows data owners to sanitize their data locally and independently (e.g., on mobile devices) before sending them to the server, and becomes a popular notion of data privacy over the decentralized Internet. It has been deployed for user data analyses in the Google Chrome web browser [5], usage data collection in the Microsoft Windows OS [6], and user data collection in the Apple iOS/MacOS [7]. Though the local privacy model has the advantage of trusting no parties (including the potential compromised/curious server), it is criticized for bringing too much noise to estimators of interests [8]. For instance, in a task of binary value summation with $n$ data owners and privacy budget $\epsilon$, the centralized model can be performed with total variance $O\left(\frac{1}{\epsilon^2}\right)$. Meanwhile, in the local model, the injected random noise is at least $O\left(\frac{n}{\epsilon^2}\right)$.

Recent studies search the middle ground between the centralized and local model, and propose to anonymize and shuffle local private messages contributed by owners [9] with semi-trusted shufflers. As Fig. 1 demonstrated, compared to the centralized model that relies on trusted curators, the shuffle model only assumes semi-trust shufflers (as the private views can be encrypted), who can be easily simulated by anonymous channels [10], secure hardware [9], cryptography shared randomness [11], or semi-trusted edge servers. Compared to the local model that every owner injects sufficient noises independently, the private views in the shuffle model could hide among anonymous messages from others, hence only needs to inject a part of noises in the local. Formal guidance of local noises injection comes from the phenomenon of *privacy amplification by shuffling* (see [8], [12], [13], [14]), which shows $n$ local $\epsilon_l$-differential-private and then shuffled messages satisfy centralized $O\left(\sqrt{\frac{e_l^{\epsilon} \log 1/\delta}{n}}\right)$-differential privacy with probability

Manuscript received 17 February 2022; revised 1 December 2022; accepted 7 February 2023. Date of publication 22 February 2023; date of current version 4 April 2023. The work of Shaowei Wang was supported in part by the National Natural Science Foundation of China under Grant 62102108, in part by the National Key Research and Development (R&D) Program Young Scientist Scheme No. 2022YFB3102400, in part by the Natural Science Foundation of Guangdong Province of China under Grant 2022A1515010061, and in part by Guangzhou Basic and Applied Basic Research Foundation under Grants 202201010194 and 622191-098. The work of Kongyang Chen was supported in part by the National Natural Science Foundation of China under Grant 61802383, in part by Research Project of Pazhou Lab for Excellent Young Scholars under Grant PZL2021KF0024, and in part by Guangzhou Basic and Applied Basic Research Foundation under Grants 202201010330 and 202201020162. This work is also supported in part by the National Natural Science Foundation of China for Joint Fund Project under Grant U1936218, in part by the National Natural Science Foundation of China under Grant 62102107, in part by the National Natural Science Foundation of China under Grant 62072132, and in part by the National Natural Science Foundation of China under Grant 62002074. Recommended for acceptance by J. Wang. *(Corresponding author: Shaowei Wang.)*

Shaowei Wang, Xuandi Luo, and Qi Chen are with the Institute of Artificial Intelligence and Blockchain, Guangzhou University, Guangzhou 511442, China (e-mail: wangsw@mail.ustc.edu.cn; 2112006166@e.gzhu.edu.cn; chenqi.math@gmail.com).

Yuqiu Qian is now an independent researcher, with partial work done at the Interactive Entertainment Group, Tencent Inc., Shenzhen, Guangdong Province 518071, China (e-mail: yuqiuqian@tencent.com).

Kongyang Chen is with the Institute of Artificial Intelligence and Blockchain, Guangzhou University, Guangzhou 511442, China, and also with Pazhou Lab, Guangzhou 510330, China (e-mail: kychen@gzhu.edu.cn).

Youwen Zhu is with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu Province 210016, China (e-mail: zhuyw@nuaa.edu.cn).

Bangzhou Xin and Wei Yang are with the Department of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230052, China (e-mail: xbw@mail.ustc.edu.cn; qubit@ustc.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPDS.2023.3247541, provided by the authors.

Digital Object Identifier 10.1109/TPDS.2023.3247541

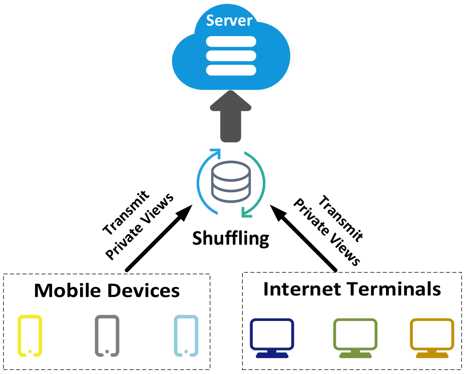


Fig. 1. Demonstration of data aggregation in the shuffle privacy model.

$1-\delta$. Consequently, for data owners with desired centralized privacy level $\epsilon_c$, their local privacy budgets can be enlarged to $\epsilon_l = \max\left\{\tilde{O}\left(\log n\epsilon_c^2\right), \epsilon_c\right\}$, which surpasses $\epsilon_c$ when population size $n$ is relatively large.

Though the population size $n$ is a critical parameter in the shuffle model, existing works on the theoretical aspect (e.g., privacy amplification lower bound [14], data utility upper bound [15]) or the application aspect (e.g., for binary summation [5], [16], distribution estimation [17], and federated learning [18], [19]) are relying on the assumption that the population size is fixed and is known at the beginning. However, in realistic data aggregation scenarios, the participating population is usually highly dynamic and uncertain. For instance, the number of mobile devices that visit a cellular base station or a wireless access point varies every day [20], the number of vehicles that pass by a location changes every hour [21]. These scenarios call for the shuffle privacy model with random population size, where the population size $N$ (i.e., the number of messages that one can hide among is uncertain) is a random variable following a distribution $P_N$ (e.g., binomial, normal, and Poisson distribution). Two naive strategies that use the existing fixed-$n$ shuffle model might be considered as solutions:

a) *Conservative Population Size:* When the population size is random, one can simply assume that the actual population size is the possible minimal $N_{min}$, and amplify privacy with $N_{min}$. However, the $N_{min}$ may be too small to provide meaningful amplification results.

b) *Expected Population Size:* One may also simply assume that the population size is the expected number $\mathbb{E}_{n\sim P_N}[n]$. However, the privacy amplification bound with $\mathbb{E}_{n\sim P_N}[n]$ might be too optimistic, and violates privacy constraints when $P_N$ is in ill forms (e.g., with extremely large variance, see Section IV-D).

In this work, we initialize the study of shuffle differential privacy with random participating population and provide formal privacy amplification results that take into account the randomness of population size. Based on the prominent *clone* analogy for privacy amplification [14] (see Section III-D) and lower bound analyses on number of clones, we derive amplification upper bounds of shuffled local private messages for population size with commonly encountered distributions: binomial, Poisson, sub-Gaussian, sub-exponential, and etc. Furthermore, we formulate the problem of adding dummies in the single-message shuffle model with a random population, and seek optimal dummy sizes therein for improving the estimation accuracy. Since the single-message shuffle model usually faces higher error barriers [22], we finally provide multi-message shuffle private protocols for random population size, which exploits distributed noises that are (almost) closed under summation. The contributions of this work can be summarized as follows:

I. We give formal privacy amplification guarantees for the shuffle privacy model with a random population, which connects the amplification bound with the tail probability of the population size distribution.

II. We derive privacy lower amplification bounds for population size with binomial, Poisson, sub-Gaussian, sub-exponential or general distributions.

III. We propose two dummy methods for improving information elicitation in the shuffle model with random population size and derive optimal size therein.

IV. We provide protocols for multi-message shuffle model with a random population, and derive noise decomposition strategies when user population is random.

V. On both real-world and synthetic datasets, we demonstrate significant improvements of our proposals over naive approaches.

The remainder of this paper is organized as follows. Section II reviews related works. Section III provides preliminary knowledge about the centralized, local, and shuffle model of differential privacy. Section IV formulates the shuffle privacy model with a random population, and gives privacy amplification results for population size with certain distributions. Section V proposes dummy methods for optimizing aggregation utility. Section VI provides secure and private protocols for the multi-message shuffle model with random population size. Section VII presents experimental results. Finally, Section VIII concludes the whole paper.

## II. Related Works

In this section, we review works on differential privacy in distributed settings, and retrospect privacy amplification in the shuffle privacy model.

### A. *Distributed Differential Privacy*

Originated from the database community for statistical queries, differential privacy [1] now becomes the *de facto* notion of privacy in distributed networking environments (e.g., in Internet services [5], mobile computing [6], and edge computing [23]). Due to its conciseness and efficiency, many works focus on the local model [4] of differential privacy, where every participant sanitizes their data locally and independently (e.g., on mobile devices or sensors). Plenty of theoretical results; [16], [17], [29] or categorical units [26], [30]), such as with Binomial noises in [16], decomposed Laplace/Geometric noises in [17], [26]. Single-message shuffle private protocols mainly exploit the privacy amplification phenomenon of local-private messages [8], as will be explained in the following parts.

TABLE I
LIST OF NOTATIONS

| Notation | Description |
|---|---|
| $n'$ | The number of all potential users (data owners) |
| $n$ | The number of participated users |
| $N$ | The variable denoting the number of participated users |
| $P_N$ | The probability distribution of $N$ |
| $\epsilon_c$ | The privacy budget in the centralized model |
| $\epsilon_l$ | The privacy budget in the local model |
| $k$ | The number of dummies |
| $l$ | The number of queries (tasks) |

### B. Shuffle Privacy Amplification

An interesting phenomenon in the shuffle model is the privacy amplification [8], which means a lower level of local privacy actually satisfies a higher level of centralized privacy after shuffling (i.e., by anonymously hiding among other users' messages). Exploiting the uniform random distribution over the binary domain, the seminal work of [8] first derived that $n$ binary randomized response messages (with local budget $\epsilon_l$) satisfies centralized $\left(\sqrt{\frac{144\epsilon_l^2\log(1/\delta)}{n}}, \delta\right)$-differential privacy. Latterly, the work of [13] shows privacy amplification also exists for the randomized response with $c$ options and gives a centralized privacy bound of $\left(\sqrt{\frac{14(\epsilon_l+c-1)\log(2/\delta)}{n-1}}, \delta\right)$. Recently, through the clone perspective, the work of [14] shows that privacy amplification actually exists for any local private mechanisms. Besides privacy amplification lower bounds, there are also several works deal with amplification lower bounds (e.g., in [12], [15]), and give $(\epsilon_l - \log n, \delta)$ as the amplification limit. It should be noted that existing works on the shuffle privacy model (including the variant: anonymous random check-in [31]), are assuming the population size that every participant could hide from is fixed. This work initializes the study of shuffle privacy model with random population size, for covering real-world scenarios where the participating population is highly dynamic and unpredictable.

## III. Preliminaries

We here retrospect the definition of differential privacy in the centralized, local, and shuffle models. Then, we provide prior knowledge about privacy amplification in the shuffle model with a fixed population and restate the clone analog [14] for deriving privacy amplification bounds. As ordering matters in the shuffle model, we use curly brackets $\{\,\}$ to denote unordered sets/multisets and use brackets $[\,]$ to denote ordered lists. We list commonly used notations in Table I.

### A. Differential Privacy

Let $\mathcal{X}$ denote the domain of every user's data, the data from $n$ users forms a dataset $D \in \mathcal{X}^n$. For any pair of datasets $D$, $D' \in \mathcal{X}^n$ that are of the same size and differ only in one element, they are called *neighboring datasets*. The centralized differential privacy with budget/level $(\epsilon, \delta)$ is as follows.

*Definition 1 (Centralized $(\epsilon, \delta)$-DP [1]).* Let $\mathcal{D}_K$ denote the output domain, a randomized mechanism $K$ satisfies $\epsilon$-differential privacy iff for any neighboring datasets $D, D' \in \mathcal{X}^n$, and any subset of outputs $\mathbf{z} \subseteq \mathcal{D}_K$, random variables $K(D)$ and $K(D')$ are $(\epsilon, \delta)$-indistinguishable. That is,

$$\Pr\left[K(D) \in \mathbf{z}\right] \leq \exp\left(\epsilon\right) \cdot \Pr\left[K(D') \in \mathbf{z}\right] + \delta.$$

The $(\epsilon, \delta)$-DP can be interpreted as satisfying pure $(\epsilon, 0)$-DP with a probability at least $1 - \delta$. This interpretation will facilitate the proofs of $(\epsilon, \delta)$-DP for both single-message and multi-message shuffle private protocols.

### B. Local Differential Privacy

Let $K$ denote a randomized mechanism for sanitizing a single user data, the definition of local DP with privacy budget $\epsilon$ is presented in Definition 2. Essentially, the local DP allows each user to sanitize their data independently, without trusting any parties (e.g., the server, database curators, other users).

*Definition 2 (Local $\epsilon$-DP [4]).* Let $\mathcal{D}_K$ denote the output domain, a randomized mechanism $K$ satisfies local $\epsilon$-differential privacy iff for any data pair $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, and any output $z \in \mathcal{D}_K$,

$$\Pr\left[K(\mathbf{x}) = z\right] \leq \exp(\epsilon) \cdot \Pr\left[K(\mathbf{x}') = z\right].$$

### C. Shuffle Differential Privacy

When semi-trusted shufflers (or anonymous channels) lie between data owners and the server, the local private views are further uniform-randomly shuffled and anonymized (see Fig. 1). Therefore, the privacy level can be amplified in the centralized perspective. Let $K$ denote the local randomizer, and $t_i = K(\mathbf{x}_i)$ denote the local $\epsilon_l$-DP view from data owner $i$, the definition of (single-message) shuffle $(\epsilon, \delta)$-DP as follows.

*Definition 3 (Single-message Shuffle $\epsilon$-DP [8], [12]).* The randomized mechanisms $K$ satisfies shuffle $(\epsilon, \delta)$-differential privacy iff the outputting unordered set/multiset $\{t_1, t_2, \ldots, t_n\}$ satisfies centralized $(\epsilon, \delta)$-DP constraints for the dataset $D = [x_1, x_2, \ldots, x_n]$.

When each data owner is allowed to sending multiple messages to the shuffler (i.e., $t_i$ is set-valued), the server only observes the union set $\bigcup_{i=1}^n t_i$. We have the definition of (multi-message) shuffle $(\epsilon, \delta)$-DP is as follows.

*Definition 4 (Multi-message Shuffle $\epsilon$-DP [12], [22]).* The randomized mechanisms $K$ satisfies shuffle $(\epsilon, \delta)$-differential privacy iff the union unordered set/multiset $\bigcup_{i=1}^n t_i$ satisfies centralized $(\epsilon, \delta)$-DP constraints for the dataset $D = [x_1, x_2, \ldots, x_n]$.

*Security Assumptions:* Following conventions in the literature, we assume the shuffler and all users are semi-honest, and the server does not collude with the shuffler or users. That is, they follow the local randomization & uniform shuffling & aggregation protocol, but may try to peek privacy information of other parties. Every user encrypts their message(s) with probabilistic encryption (e.g., additive Homomorphic encryption [32]) before sending them to the shuffler. Therefore, the shuffler can not infer raw values of messages, while the statistician (e.g., the aggregator on the server side) can still decrypt shuffled messages. In the shuffle privacy model, the privacy guarantee of the message(s) from one user relies on other users. For simplicity, we assume that all users follow the protocols and does not collude with other parties, as straight-forward strategies can deal with a limited fraction of malicious/adversarial/corrupted users [33].

### D. Privacy Amplification of the Single-Message Shuffle Model

The seminal work of [12] considers $n$ local $\epsilon_l$-DP messages $T = \{t_1, \ldots, t_n\}$ $(t_i \in \{0,1\})$ processed by binary randomized response mechanism [34], [35]. After shuffling, the server only observes the frequencies $F(z) = \#\{t_i \mid for\ i \in [1, \ldots, n]\ and\ t_i = z\}$ supported in $z \in \{0,1\}$. Therefore, in order to prove that $T$ satisfies centralized $(\epsilon_c, \delta)$-DP, [12] find that it is enough to show that the $F$ satisfies centralized $(\epsilon_c, \delta)$-DP. Since about $\frac{2(n-1)}{e^{\epsilon}+1}$ users responded uniform randomly, they contributed $\mathbf{B}\left(\frac{2(n-1)}{e^{\epsilon}+1}, 0.5\right)$ Binomial random noises to each frequency, it is guaranteed that the $F$ satisfies $\left(\sqrt{144\ln(1/\delta)\frac{e^{\epsilon_l}+1}{n}}, \delta\right)$-DP. Following-up works have improved the amplification bound to $\left(\sqrt{14\ln(2/\delta)\frac{e^{\epsilon_l}+1}{n-1}}\right.$ and given bounds also for multinomial randomized response [13].

Recently, Feldman et al. [14] shows that a similar privacy amplification effect holds for arbitrary local private randomizers beyond randomized response, they introduce the concept of *clone* for deriving amplification bounds. Considering the neighboring datasets differing at the user 1, it uses the observation that for any $x_i$ and $x_1 \in \{a, b\}$:

$$\mathcal{M}(x_i) = \begin{cases} \mathcal{M}(a), & \text{with probability } \frac{1}{2\exp(\epsilon_l)}; \\ \mathcal{M}(b), & \text{with probability } \frac{1}{2\exp(\epsilon_l)}; \\ \mathcal{W}(x_i), & \text{else.} \end{cases} \tag{1}$$

holds for a certain randomization process $\mathcal{W}$. That is, every private view of $x_i$ is a clone of $x_1 = a$ or $x_1 = b$ with probability $\frac{1}{\exp(\epsilon_l)}$, hence the true value of $x_1$ could hide among these clones.

Based on the concept of clone, for proving the shuffled messages $T$ satisfy centralized $(\epsilon_c, \delta)$-DP, it is sufficient to show that the frequencies of $\mathcal{M}(a)$ and $\mathcal{M}(b)$ satisfy centralized $(\epsilon_c, \delta)$-DP, which can be analogously induced as in Lemma 1 [14].

*Lemma 1 (Shuffle Privacy via Clone Analogy [14]).* Consider the process that we first sample $C \sim Binomial\left(n-1, \frac{1}{e^{\epsilon_l}}\right)$, then $A \sim Binomial\left(C, \frac{1}{2}\right)$, and sample $B \sim Bernoulli\left(\frac{e^{\epsilon_l}}{e^{\epsilon_l}+1}\right)$, then the differential private constraint that $T_a = \{\mathcal{M}(a), \mathcal{M}(x_1), \ldots, \mathcal{M}(x_n)\}$ and $T_b = \{\mathcal{M}(b), \mathcal{M}(x_1), \ldots, \mathcal{M}(x_n)\}$ are $(\epsilon_c, \delta)$-indistinguishable is equivalent to: random variables $P = (A+B, C-A+1-B)$ and $Q = (A+1-B, C-A+B)$ are $(\epsilon_c, \delta)$-indistinguishable.

It is also observed that $\mathcal{M}(b)$ itself is a clone of $\mathcal{M}(a)$ with probability $\frac{1}{e^{\epsilon_l}}$, and the Bernoulli variable $B$ records the event of being a clone of $\mathcal{M}(b)$ (or$\mathcal{M}(a)$) for the user 1 with $x_1 = a$ (or $x_1 = b$). The variable $A$ (or $C - A$) records the number of clones $\mathcal{M}(a)$ (or $\mathcal{M}(b)$) from users $[2 : n']$. With a careful distinguishability bounding on the counts $(A + B, C - A + 1 - B)$ and $(A + 1 - B, C - A + B)$, Feldman et al. [14] show that they are $(\epsilon_c, \delta)$-indistinguishable with $\epsilon_c \le \log\left(1 + \frac{e^{\epsilon_l}-1}{e^{\epsilon_l}+1}\left(8\sqrt{\frac{e^{\epsilon_l}\log(4/\delta)}{n}} + \frac{8e^{\epsilon_l}}{n}\right)\right)$. That is, for any participant that hides among $n$ local $\epsilon_l$-DP and shuffled messages, their centralized DP level is guaranteed to be $\left(\log\left(1 + \frac{e^{\epsilon_l}-1}{e^{\epsilon_l}+1}\left(8\sqrt{\frac{e^{\epsilon_l}\log(4/\delta)}{n}} + \frac{8e^{\epsilon_l}}{n}\right)\right), \delta\right)$. When $n$ is relatively large, the amplified centralized level is more stringent than the local privacy level $\epsilon_l$.

## IV. Single-Message Shuffle Model With Random Population

As opposed to the case in previous studies [8], [13], [14] that every user participates with certain, this section formulates the shuffle private model with a random participating population. We first simulate the participating choices of $n'$ possible users as random variables $M \in \{0,1\}^{n'}$, and then give general privacy amplification bounds regarding the distribution (or statistics like mean and variance) of population size $|M|$. Specifically for binomial, Poisson, sub-Gaussian, and sub-exponential distributed population size, we derive tighter bounds.

### A. Problem Formulation

Let 1 indicate the event of participating and 0 indicate the event of not participating, the participation decisions of $n'$ potential users can be represented as a masking vector $m \in \{0,1\}^{n'}$. To model sophisticated decisions of $n'$ users, we use $\mathcal{P}_M$ to denote the probability distribution of the masking vector (variable) $M$.

When every potential user participates as in the original shuffle privacy model, the $\mathcal{P}_M$ is:

$$\mathcal{P}_M(m) = \begin{cases} 1, & \text{if } m = [1, 1, \ldots, 1]; \\ 0, & \text{else.} \end{cases}$$

When the shuffler samples $n$ users for participating (e.g., for batch stochastic gradient descent in federated learning [18], [19]), here the $\mathcal{P}_M$ is:

$$\mathcal{P}_M(m) = \begin{cases} 1/\binom{n'}{n}, & \text{if } |m| = n; \\ 0, & \text{else.} \end{cases}$$

The previous two cases have fixed population sizes, another slightly more complicated case is that every potential user decides to participate independently with rate $\alpha$, then the $\mathcal{P}_M$ is:

$$\mathcal{P}_M(x) = \alpha^{|x|}(1-\alpha)^{n'-|x|}.$$

In real-world aggregation scenarios, the participating distribution $\mathcal{P}_M$ is usually in more complex forms (e.g., when decisions are correlated).

From an incoming user's perspective, w.l.o.g. denoted as the user 1, the conditional probability distribution $\mathcal{P}_{M \in 1 \times \{0,1\}^{n'-1}}$ is of more interest. Under this participating model, similar to the original shuffle differential privacy Definition 3, we can define the (single-message) shuffle $(\epsilon, \delta)$-DP [8] with random population for the user 1 as follows, which takes into account both the randomness of $K$ and the randomness of participating population size.

*Definition 5 (Shuffle $(\epsilon, \delta)$-DP [8] with random population).* The randomized mechanism $K$ satisfies shuffle $(\epsilon, \delta)$-differential privacy iff when $m \sim \mathcal{P}_{M \in 1 \times \{0,1\}^{n'-1}}$, the outputting unordered set $T = \{t_i\}_{i \in m}$ (here $t_i = K(x_i)$) are centralized $(\epsilon, \delta)$-indistinguishable for any possible neighboring datasets: $D = [x_1 = a, x_2, \ldots, x_{n'}]$ and $D' = [x_1 = b, x_2, \ldots, x_{n'}]$ $(a, b \in \mathcal{X})$.

Apparently, the probability distribution of the output $T$ could be seen as a mixture distribution: $\sum_{m \in 1 \times \{0,1\}^{n'-1}} \Pr[M = m] \cdot \Pr[T|m]$. Here every $\Pr[T|m]$ is an instance of the original shuffle model. Recall that the privacy amplification bound depends only on the population size (as in Lemma 1). Let $\mathcal{P}_N$ denote the population size distribution: $\mathcal{P}_N(n) = \sum_{m \in 1 \times \{0,1\}^{n'-1} \textit{ and } |m| = n} \Pr[M = m]$, we then have the following lemma for shuffle model with random population. This lemma is a straight-forward extension of Lemma 1.

*Lemma 2 (Shuffle Privacy via Clone Analogy with Random Population).* Consider the process that we first sample $N \sim \mathcal{P}_N$, then sample $C \sim Binomial\left(N-1, \frac{1}{e^{\epsilon_l}}\right)$ and sample $A \sim Binomial\left(C, \frac{1}{2}\right)$, and sample $B \sim Bernoulli\left(\frac{e^{\epsilon_l}}{e^{\epsilon_l}+1}\right)$, then the differential private constraint that $T_a = \{\mathcal{M}(a), \mathcal{M}(x_1), \ldots, \mathcal{M}(x_n)\}$ and $T_b = \{\mathcal{M}(b), \mathcal{M}(x_1), \ldots, \mathcal{M}(x_n)\}$ are $(\epsilon_c, \delta)$-indistinguishable is equivalent to: random variables $P = (A + B, C - A + 1 - B)$ and $Q = (A + 1 - B, C - A + B)$ are $(\epsilon_c, \delta)$-indistinguishable.

The Bernoulli variable $B$ records the event of being a clone of $\mathcal{M}(b)$ (or$\mathcal{M}(a)$) for the user 1 with $x_1 = a$ (or $x_1 = b$). The variable $A$ or $C - A$ records the number of clones $\mathcal{W}(a)$ or $\mathcal{W}(b)$ from users $[2 : n']$ respectively. Building upon the analogy, the problem of lower bounding privacy amplification with a random population is now simplified to analyzing the probability distributions of $P$ and $Q$. In the following subsections, we separately consider cases when $\mathcal{P}_{N-1}$ follows Binomial, Poisson, sub-Gaussian (see Appendix E, available online), sub-exponential (see Appendix A, available in the online supplemental material), or general variance-bounded distributions.

### *B. Privacy Amplification With Binomial Distribution*

Recall that when users $[2 : n']$ decide to participate independently with a Bernoulli success rate $\alpha$, we have $N - 1 \sim Binomial\,(n' - 1, \alpha)$. In Theorem 1, observing that $C \sim Binomial\left(n' - 1, \frac{\alpha}{e^{\epsilon_l}}\right)$, we show that the $P = (A + B, C - A + 1 - B)$ and $Q = (A + 1 - B, C - A + B)$ are $\left(log\left(1 + \frac{e^{\epsilon_l}-1}{e^{\epsilon_l}+1} \cdot \frac{2\sqrt{\Omega/2 \log(4/\delta)}+1}{\Omega/2 - \sqrt{\Omega/2 \log(4/\delta)}}\right), \delta\right)$-indistinguishable. The variable $\Omega$ generally grows with $\alpha$ $\left(\text{when}\alpha \geq \frac{e^{\epsilon_l}}{n'}\sqrt{\frac{3\log(2/\delta)}{4}}\right)$, thus a higher Bernoulli success rate usually means a more significant privacy amplification.

*Theorem 1.* Consider the process that we first sample $N - 1 \sim Binomial\,(n' - 1, \alpha)$, then sample $C \sim Binomial\left(n - 1, \frac{1}{e^{\epsilon_l}}\right)$, sample $A \sim Binomial\left(C, \frac{1}{2}\right)$, and sample $B \sim Bernoulli\left(\frac{e^{\epsilon_l}}{e^{\epsilon_l}+1}\right)$, then random variables $P = (A + B, C - A + 1 - B)$ and $Q = (A + 1 - B, C - A + B)$ are $(\epsilon_c, \delta)$-indistinguishable for $\epsilon_c = \log\left(1 + \frac{e^{\epsilon_l}-1}{e^{\epsilon_l}+1} \cdot \frac{2\sqrt{\Omega/2 \log(4/\delta)}+1}{\Omega/2 - \sqrt{\Omega/2 \log(4/\delta)}}\right)$, when $\Omega = \frac{n'\alpha}{e^{\epsilon_l}} - \sqrt{3\frac{n'\alpha}{e^{\epsilon_l}} \log(4/\delta)} > 2\log(4/\delta)$. Specifically, when $n' \frac{\alpha}{e^{\epsilon_l}} \geq 3\log(4/\delta)$, we can also simplify it as:

$$\epsilon_c = \log\left(1 + \frac{e^{\epsilon_l} - 1}{e^{\epsilon_l} + 1} \cdot \left(8\sqrt{\frac{e^{\epsilon_l}\log(4/\delta)}{\alpha n'}} + \frac{8e^{\epsilon_l}}{\alpha n'}\right)\right).$$

*Proof.* When $N - 1 \sim Binomial\,(n' - 1, \alpha)$ and $C \sim Binomial\left(N - 1, \frac{1}{e^{\epsilon_l}}\right)$, since $Bernoulli(\alpha) \cdot Bernoulli\left(\frac{1}{e^{\epsilon_l}}\right)$ is equivalent to $Bernoulli\left(\frac{\alpha}{e^{\epsilon_l}}\right)$, we have the number of clones $C \sim Binomial\left(n' - 1, \frac{\alpha}{e^{\epsilon_l}}\right)$. Hence, according to the Chernoff bound, with probability at most $\delta/2$, we have:

$$C < \frac{n'\alpha}{e^{\epsilon_l}} - \sqrt{\frac{3n'\alpha}{e^{\epsilon_l}}\log(4/\delta)}.$$

Applying Hoeffding's inequality on the Binomial variable $A$, with probability at most $\delta/2$, we have:

$$|A - C/2| > \sqrt{C\log(4/\delta)/2}.$$

Then, according to the union bound of probabilities, with probability at least $1 - \delta$, both $C\frac{(n'-1)\alpha}{e^{\epsilon_l}} - \sqrt{\frac{3(n'-1)\alpha}{e^{\epsilon_l}}\log(4/\delta)}$ and $|A - C/2| \leq \sqrt{C\log(4/\delta)/2}$ hold.

Define $P' = (A + 1, C - A)$ and $Q' = (A, C - A + 1)$, since the following equality always hold [14]: $\Pr\left[P' = (a, b)\right] = \Pr\left[C = a + b + 1\right] \cdot \Pr\left[A = a - 1 \mid C = a + b + 1\right] = \Pr\left[C = a + b + 1\right] \cdot \Pr\left[A = a \mid C = a + b + 1\right] \cdot \frac{a}{b} = \Pr\left[Q' = (a, b)\right] \cdot \frac{a}{b}$, we then have (with probability $1 - \delta$):

$$\begin{aligned}\frac{\Pr[P = (a, b)]}{\Pr[Q = (a, b)]} &= \frac{a}{b} = \frac{A + 1}{C - A} \\ &\leq \frac{C/2 + \sqrt{C/2\log(4/\delta)} + 1}{C/2 - \sqrt{C/2\log(4/\delta)}}. \qquad (2)\end{aligned}$$

Since $\frac{C/2 + \sqrt{C/2\log(4/\delta)}+1}{C/2 - \sqrt{C/2\log(4/\delta)}}$ deceases with $C$ when $C \geq 1 + \frac{1 + \sqrt{1 + 2\log(4/\delta)}}{\log(4/\delta)}$ and $C > 2\log(4/\delta)$, we have the bound by replacing $C$ as $\Omega = \frac{n'\alpha}{e^{\epsilon_l}} - \sqrt{\frac{3n'\alpha}{e^{\epsilon_l}}\log(4/\delta)}$. Hence $P'$ and $Q'$ are $\left(\log\left(1 + \frac{e^{\epsilon_l}-1}{e^{\epsilon_l}+1} \cdot \frac{2\sqrt{\Omega/2\log(4/\delta)}+1}{\Omega/2 - \sqrt{\Omega/2\log(4/\delta)}}\right), \delta\right)$-indistinguishable. Now consider $P = (A + B, C - A + 1 - B)$ and $Q = (A + 1 - B, C - A + B)$, since $B \sim Bernoulli\left(\frac{e^{\epsilon_l}}{e^{\epsilon_l}+1}\right)$, according to privacy accountant with sub-sampling [36], we have $P$ and $Q$ are $\left(\log\left(1 + \frac{e^{\epsilon_l}-1}{e^{\epsilon_l}+1} \cdot \frac{2\sqrt{\Omega/2\log(4/\delta)}+1}{\Omega/2 - \sqrt{\Omega/2\log(4/\delta)}}\right), \delta\right)$-indistinguishable. □

### C. Privacy Amplification With Poisson Distribution

Poisson distribution is widely used for modeling the number of events occurring within a fixed interval of time or region, and naturally fits shuffle private aggregation scenarios where shufflers (e.g., edge servers) publish messages in a daily or hourly manner. When the number of participating population (except user 1) follows the Poisson distribution with mean value $\lambda$ (i.e., $N-1 \sim Poisson(\lambda)$), we give an indistinguishable level of the $P=(A+B, C-A+1-B)$ and $Q=(A+1-B, C-A+B)$ in Theorem 2 (see Appendix B, available in the online supplemental material for proof). Since the variable $\Omega$ grows with $\lambda$ $\left(\text{when}\lambda \geq e^{\epsilon_l}\sqrt{\frac{\log(2/\delta)}{2}}\right)$, we have the privacy amplification effects grow with the Poisson mean.

*Theorem 2.* Consider the process that we first sample $N-1 \sim Poisson(\lambda)$, then sample $C \sim Binomial\left(n-1, \frac{1}{e^{\epsilon_l}}\right)$, sample $A \sim Binomial\left(C, \frac{1}{2}\right)$, and sample $B \sim Bernoulli\left(\frac{e^{\epsilon_l}}{e^{\epsilon_l}+1}\right)$, then random variables $P=(A+B, C-A+1-B)$ and $Q=(A+1-B, C-A+B)$ are $(\epsilon_c, \delta)$-indistinguishable for:

$$\epsilon_c = \log\left(1+\frac{e^{\epsilon_l}-1}{e^{\epsilon_l}+1}\cdot\frac{2\sqrt{\Omega/2\log(4/\delta)}+1}{\Omega/2-\sqrt{\Omega/2\log(4/\delta)}}\right),$$

when $\Omega = \frac{\lambda}{e^{\epsilon_l}} - \sqrt{\frac{2\lambda}{e^{\epsilon_l}}\log(2/\delta)} > 2\log(4/\delta)$.

### D. Distributions With Bounded Mean and Variance

In some scenarios, we may not know the exact or rough shape of the probability distribution of $N-1$, and only possess a few statistics (e.g., mean, variance) about $N-1$. Therefore, this subsection aims to provide privacy amplification lower bounds for random population size with bounded mean and variance.

As the key ingredient to provide amplification lower bound is tail bounding the number of clones $C$, we now first derive the mean $\mu_c$ and variance $\mu_c$ of variable $C$. Let $\mu$ and $\tau$ denote the mean and variable of $N-1 \sim \mathcal{P}_{N-1}$, since $\Pr[C=c] = \sum_{n-1=0}^{+\infty} \mathcal{P}_N(n-1)\binom{n-1}{c}\frac{1}{e^{c\epsilon_l}}\left(\frac{e^{\epsilon_l}-1}{e^{\epsilon_l}}\right)^{n-1-c}$, we have $\mu_c$ as $\frac{\mu}{e^{\epsilon_l}}$, similarly we have $\tau_c^2+\mu_c^2$ as:

$$\begin{aligned}
\mathbb{E}[C^2] &= \sum_{c=0}^{+\infty} c^2\cdot\Pr[C=c] \\
&= \sum_{c=0}^{+\infty} c^2\cdot \sum_{n-1=0}^{+\infty} \mathcal{P}_N(n-1)\binom{n-1}{c}\frac{1}{e^{c\epsilon_l}}\left(\frac{e^{\epsilon_l}-1}{e^{\epsilon_l}}\right)^{n-1-c} \\
&= \sum_{n-1=0}^{+\infty} \mathcal{P}_N(n-1)\sum_{c=0}^{+\infty} c^2\cdot\binom{n-1}{c}\frac{1}{e^{c\epsilon_l}}\left(\frac{e^{\epsilon_l}-1}{e^{\epsilon_l}}\right)^{n-1-c} \\
&= \sum_{n-1=0}^{+\infty} \mathcal{P}_N(n-1)\frac{(n-1)(e^{\epsilon_l}-1)+(n-1)^2}{e^{2\epsilon_l}} \\
&= \frac{\mu(e^{\epsilon_l}-1)+\tau^2+\mu^2}{e^{2\epsilon_l}}.
\end{aligned}$$

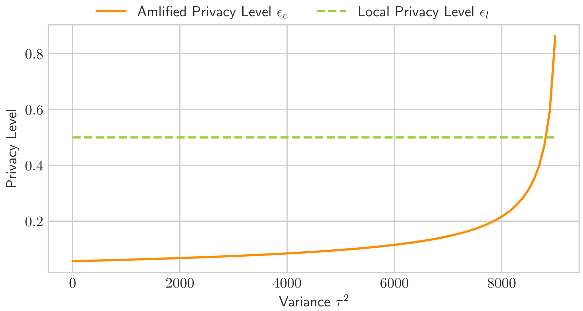


Fig. 2. Privacy amplification under random population with vary variance.

That is, $\tau_c = \sqrt{\frac{\mu_{n-1}(e^{\epsilon_l}-1)+\tau_{n-1}^2+\mu_{n-1}^2}{e^{2\epsilon_l}} - \frac{\mu_{n-1}^2}{e^{2\epsilon_l}}} = \sqrt{\frac{\mu_{n-1}(e^{\epsilon_l}-1)+\tau_{n-1}^2}{e^{2\epsilon_l}}}$.

Given $\mu_c$ and $\tau_c$, we now give privacy amplification lower bounds for random population size with bounded mean and variance in Theorem 3.

*Theorem 3.* For a distribution $\mathcal{P}_{N-1}$ with mean value $\mu$ and variance $\tau^2$, consider the process that we first sample $N-1 \sim \mathcal{P}_{N-1}$, then sample $C \sim Binomial\left(N-1, \frac{1}{e^{\epsilon_l}}\right)$, sample $A \sim Binomial\left(C, \frac{1}{2}\right)$, and sample $B \sim Bernoulli\left(\frac{e^{\epsilon_l}}{e^{\epsilon_l}+1}\right)$, then random variables $P=(A+B, C-A+1-B)$ and $Q=(A+1-B, C-A+B)$ are $(\epsilon_c, \delta)$-indistinguishable with:

$$\epsilon_c = \log\left(1+\frac{e^{\epsilon_l}-1}{e^{\epsilon_l}+1}\cdot\frac{2\sqrt{\Omega/2\log(4/\delta)}+1}{\Omega/2-\sqrt{\Omega/2\log(4/\delta)}}\right),$$

when $\Omega = \frac{\mu}{e^{\epsilon_l}} - \sqrt{\frac{\mu(e^{\epsilon_l}-1)+\tau^2}{0.5\delta e^{2\epsilon_l}}} > 2\log(4/\delta)$.

*Proof.* According to the Markov's inequality, for $\Delta > 0$, we have:

$$\begin{aligned}
&\Pr[c < \mu_c - \Delta] = \Pr[\mu_c - c > \Delta] \\
&= \Pr[(\mu_c - c)^2 > \Delta^2] \leq \frac{\mathbb{E}[(\mu_c - c)^2]}{\Delta^2} \\
&\leq \frac{\mu(e^{\epsilon_l}-1)+\tau^2}{l^2 e^{2\epsilon_l}\Delta^2}.
\end{aligned}$$

Therefore, with probability at least $1-\delta$, both $C \geq \frac{\mu}{e^{\epsilon_l}} - \sqrt{\frac{\mu(e^{\epsilon_l}-1)+\tau^2}{0.5\delta e^{2\epsilon_l}}}$ and $|A - C/2| \leq \sqrt{C\log(4/\delta)/2}$ hold. Hence when $\Omega = \frac{\mu}{e^{\epsilon_l}} - \sqrt{\frac{\mu(e^{\epsilon_l}-1)+\tau^2}{0.5\delta e^{2\epsilon_l}}} > 2\log(4/\delta)$, we have the bound. □

For better illustrate the effect of the variance, in Fig. 2, we depict the amplified privacy level regarding $\tau^2$ given mean value $\mu = 5000$, $\delta = 0.0001$, and local privacy budget $\epsilon_l = 0.5$. Compared to the classical shuffle privacy model with fixed $N-1 \equiv 5000$ (i.e., when $\tau^2 = 0$), the amplification bound gets large when the variance gets large. Specifically, when $\tau^2$ is larger than 8700, the amplification phenomenon no longer exists (i.e., $\epsilon_c \geq \epsilon_l$).

## V. Dummy Strategies

As the privacy amplification bounds highly rely on the lower tail probabilities of the population size $N-1$ and the number of clones $C$, one natural way to reduce the tail probability is hence adding dummy messages [30], [31], [37]. For example, if the shuffler (or users) adds $k$ local $\epsilon_l$-private views to the message pool, then with probability 1, we have the population size $N-1 \geq k$. On the other hand, the dummy messages added to the message pool hurt the utility of true views. In this section, we provide privacy amplification bounds with dummy messages, and derive optimal dummy size for minimizing the final estimation error (e.g., the mean squared error of binary distribution estimation).

### A. Two Dummy Methods

There are mainly two methods for adding dummy messages, one is non-adaptive and the other is adaptive. In the non-adaptive method, the shuffler (or other third parties) adds $k$ private messages to the pool in advance, without peeking the number of normal messages; while in the adaptive method, the shuffler first count the number of messages in the pool, if the number is below $k$ then padding it to $k$, otherwise no dummy messages are added.

Let $N^*-1$ denote the number of population size after adding dummy messages, and $\mathcal{P}_{N^*-1}$ denote the probability distribution of $N^*-1$, the relation between $\mathcal{P}_{N^*-1}$ with $k$-non-adaptive dummies and the true population distribution $\mathcal{P}_{N-1}$ is:

$$\mathcal{P}_{N^*-1}(n-1) = \begin{cases} 0, & \text{if } n-1 < k; \\ \mathcal{P}_{N-1}(n-1-k), & \text{else.} \end{cases} \tag{3}$$

The $\mathcal{P}_{N^*}$ with $k$-adaptive dummies is then:

$$\mathcal{P}_{N^*-1}(n-1) = \begin{cases} 0, & \text{if } n-1 < k; \\ \sum_{i=0}^{k} \mathcal{P}_{N-1}(i), & \text{if } n-1 = k; \\ \mathcal{P}_{N-1}(n-1), & \text{else.} \end{cases} \tag{4}$$

Since the adaptive dummy method simply reshapes the left tail bounds of the number of clones $C$, the corresponding privacy amplification effect is quite similar to the one without dummy, we defer the analyses on amplification bounds to Appendix C, available in the online supplemental material.

### B. Privacy Amplification With Non-Adaptive Dummies

In general, after adding dummies (i.e., $N^*-1 \sim \mathcal{P}_{N*-1}$), these cases still lie in the framework of the shuffle privacy model with a random population, except the probability distributions of the population size are changed. We proceed to provide tighter privacy amplification bounds for the random population with $k$-non-adaptive dummies. Since now $N^* = N + k$, in Theorem 4, we provide amplification results for true population size with Binomial, Poisson, sub-Gaussian and general distributions separately (see Appendix D, available in the online supplemental material for proof). As the number of dummies $k$ grows, the privacy amplification effect grows.

*Theorem 4.* Given distribution $\mathcal{P}_{N-1}$ and $\mathcal{P}_{N^*-1}$ as in (3), consider the process that we first sample $N^*-1 \sim \mathcal{P}_{N^*-1}$, then sample $C \sim Binomial\left(N^*-1, \frac{1}{e^{\epsilon_l}}\right)$, sample $A \sim Binomial\left(C, \frac{1}{2}\right)$, and sample $B \sim Bernoulli\left(\frac{e^{\epsilon_l}}{e^{\epsilon_l}+1}\right)$, then random variables $P = (A+B, C-A+1-B)$ and $Q = (A+1-B, C-A+B)$ are $\left(\log\left(1+\frac{e^{\epsilon_l}-1}{e^{\epsilon_l}+1}\cdot\frac{2\sqrt{\Omega/2\log(4/\delta)}+1}{\Omega/2-\sqrt{\Omega/2\log(4/\delta)}}\right), \delta\right)$-indistinguishable (when$\Omega > 2\log(4/\delta)$):

a). If $\mathcal{P}_{N-1}$ is $Binomial(n'-1, \alpha)$, we have:

$$\Omega = \frac{n'\alpha + k}{e^{\epsilon_l}} - \sqrt{3\frac{n'\alpha+k}{e^{\epsilon_l}}\log(4/\delta)};$$

b). If $\mathcal{P}_{N-1}$ is $Poisson(\lambda)$, we have:

$$\Omega = \frac{\lambda + k}{e^{\epsilon_l}} - \sqrt{\frac{4\lambda+2k}{e^{\epsilon_l}}\log(2/\delta)};$$

c). If $\mathcal{P}_{N-1}$ is a $\tau$-sub-Gaussian distribution with mean value $\mu$, we have $\Omega$ as:

$$\frac{k+\mu-\sqrt{\tau^2\log(4/\delta)}}{e^{\epsilon_l}} - \sqrt{\frac{k+\mu-\sqrt{\tau^2\log(4/\delta)}}{0.5e^{\epsilon_l}}\log(2/\delta)};$$

d). If $\mathcal{P}_{N-1}$ is a $(\alpha,\beta)$-sub-exponential distribution with mean value $\mu$, we have $\Omega$ as:

$$\frac{k+\mu-\frac{\log(4\alpha/\delta)}{\beta}}{e^{\epsilon_l}} - \sqrt{k+\mu-\frac{\log(4\alpha/\delta)}{\beta}0.5e^{\epsilon_l}\log(2/\delta)};$$

e). If $\mathcal{P}_{N-1}$ is a probability distribution with mean $\mu$ and variance $\tau^2$, we have:

$$\Omega = \frac{k+\mu}{e^{\epsilon_l}} - \sqrt{\frac{(k+\mu)(e^{\epsilon_l}-1)+\tau^2}{0.5\delta e^{2\epsilon_l}}}.$$

### C. Optimal Dummy Size

Though adding more dummies means every participating user could adopt a higher local budget, the dummies are contributing noises to the estimator. We now seek for the optimal choice of $k$ regarding the estimation error. Without loss of generality, we consider one of the most fundamental estimation tasks: binary distribution estimation, where the binary randomized response [34] is a universally optimal local private mechanism [38]. In the binary randomized response under local privacy budget $\epsilon_l$, given an input $x_i \in \{0,1\}$, the randomizer answers truthfully with a limited probability $\frac{e^{\epsilon_l}}{e^{\epsilon_l}+1}$ as:

$$\mathcal{M}(x_i) = \begin{cases} x_i, & \text{with probability } \frac{e^{\epsilon_l}}{e^{\epsilon_l}+1}; \\ 1-x_i, & \text{otherwise.} \end{cases} \tag{5}$$

An unbiased estimator $\hat{x}_i$ of $x_i$ is $\frac{t_i-1}{e^{\epsilon_l}-1}$. Suppose that every $x_i \in \{0,1\}$ is an i.i.d. sample from the (unknown) distribution $\mathcal{P}_x$, and there are total $n = |m|$ true participating users utilizing binary randomized response. The mean squared error (MSE) of the estimator $\hat{\mathcal{P}}_x(1) = \frac{\sum_{i\in m}\hat{x}_i}{|m|}$ is:

$$\mathbb{E}\left[|\hat{\mathcal{P}}_x - \mathcal{P}_x|_2^2 \mid |m|\right] = \frac{2e^{\epsilon_l}}{|m|(e^{\epsilon_l}-1)^2} + \frac{1-\mathcal{P}_x^2(0))-\mathcal{P}_x^2(1))}{|m|}.$$

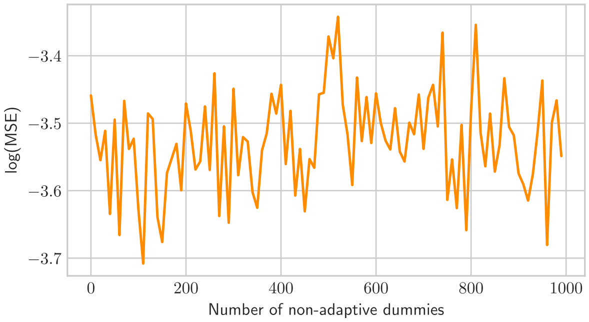


Fig. 3. Mean squared error results with vary number of non-adaptive dummies ($\epsilon_c = 0.05$, $\delta = 0.00001$, $n' = 5001$ and Bernoulli participate rate $\alpha = 0.2$, repeated 1000 times). As the experimental results showed, setting up optimal $k$ could reduce about 20% error, compared to the no dummy case (i.e., $k = 0$).

Therefore, the expected error $\mathbb{E}\left[\|\hat{\mathcal{P}}_x - \mathcal{P}_x\|_2^2\right]$ of the estimator with random participating distribution $\mathcal{P}_M$ is then:

$$\left(1 - \|\mathcal{P}_x\|_2^2 + \frac{2e^{\epsilon_l}}{(e^{\epsilon_l} - 1)^2}\right) \sum_{n'=1}^{+\infty} \frac{\mathcal{P}_{|M|}(n')}{n'}.$$

*1) Non-Adaptive Dummies:* With the $k$-non-adaptive dummy method, the true distribution of these dummies is non-private and public, but their private views also contribute noise to the estimator of $\mathcal{P}_x(1)$, and the expected error $\mathbb{E}[\|\hat{\mathcal{P}}_x - \mathcal{P}_x\|_2^2]$ becomes:

$$\sum_{n'=1}^{+\infty} \frac{\mathcal{P}_{|M|}(n')}{n'} \left(1 - \|\mathcal{P}_x\|_2^2 + \frac{2e^{\epsilon_l} + 2e^{\epsilon_l} k/n'}{(e^{\epsilon_l} - 1)^2}\right).$$

Consequently, given centralized privacy budget $\epsilon_c$, and let $\epsilon_l = A(k)$ denote the amplified local budget with $k$-non-adaptive dummies (guided by Theorem 4), the problem of finding optimal dummy size $k$ is:

$$\arg\min_{k=0}^{\infty} \sum_{n'=1}^{+\infty} \frac{\mathcal{P}_{|M|}(n')}{n'} \left(1 - \|\mathcal{P}_x\|_2^2 + \frac{2e^{A(k)} + 2e^{A(k)} k/n'}{(e^{A(k)} - 1)^2}\right),$$

which could be simplified as:

$$\arg\min_{k=0}^{\infty} \frac{2e^{A(k)}(\mathbb{E}[1/n'] + k\mathbb{E}[1/(n')^2])}{(e^{A(k)} - 1)^2}). \qquad (6)$$

In most cases, the maximal-allowed local budget $\epsilon_l = A(k)$ is an increasing function of $k$, by iterating $k$ over $[0 : n']$ with an appropriate step size (e.g., $\sqrt{n'}$), the optimization problem can be solved efficiently. For illustration, in Fig. 3, we plot the MSE w.r.t. the non-adaptive dummy size $k$ with $N - 1 \sim Binomial(5000, 0.2)$ and centralized budget $\epsilon_c = 0.05$.

*2) Adaptive Dummies:* With the $k$-adaptive dummy method, given the masking vector $m$ of participating population, the mean squared error $\mathbb{E}\left[|\hat{\mathcal{P}}_x - \mathcal{P}_x|_2^2 \mid |m|\right]$ is:

$$\frac{2e^{\epsilon_l} + 2\max\{k - |m|, 0\}/|m|e^{\epsilon_l}}{|m|(e^{\epsilon_l} - 1)^2} + \frac{1 - \|\mathcal{P}_x\|_2^2}{|m|},$$

and the expected error $\mathbb{E}\left[|\hat{\mathcal{P}}_x - \mathcal{P}_x|_2^2\right]$ is:

$$\left(1 - \|\mathcal{P}_x\|_2^2\right) \mathbb{E}\left[1/n'\right] + \frac{2e^{\epsilon_l}\left(\mathbb{E}\left[1/n'\right] + \mathbb{E}\left[\max\{k - n,' 0\}/(n')^2\right]\right)}{(e^{\epsilon_l} - 1)^2}.$$

The problem of finding optimal dummy size $k$ is then:

$$\arg\min_{k=0}^{\infty} \frac{2e^{A(k)}\left(\mathbb{E}[1/n'] + k\mathbb{E}\left[\max\{k - n,' 0\}/(n')^2\right]\right)}{(e^{A(k)} - 1)^2}). \qquad (7)$$

## VI. Multi-Message Shuffle Private Protocols

As another major stream of shuffle differential privacy, the multi-message shuffle model [22] allows every user to publish multiple messages. The definition of multi-message shuffle $(\epsilon, \delta)$-DP is identical to Definition 5 if we replace $T = \{t_i\}_{i \in m}$ with $T = \bigcup_{i \in m} t_i$. Compared to the single-message shuffle model in previous sections, it has the potential to achieve accuracy close to the centralized model [15]. One common approach to multi-message shuffle privacy is by adding random noises in a distributed way, such as decomposing discrete Laplace/Geometric noises into $n$ parts and every user contributes one part [17], [29]. However, existing noise decomposition strategies need the knowledge of population size $n$ in advance, so that every user could set up the right parameter of local noises, such as the Gamma decomposition of Laplace noises [39] and the Pólya decomposition of Geometric noises [26]. In this section, we study noise decomposition & dummy strategies when the number of participating users $n$ is random.

### A. Noise Decomposition With Random Population

*1) Direct Approach Via Union Bound:* Recall that the $(\epsilon, \delta)$-DP could be interpreted as violating $(\epsilon, 0)$-DP with probability at most $\delta$. When the participating population $n$ is random, in order to utilize existing multi-message privatization protocols designed for fixed $n$, we can split the violating probability $\delta$ into two parts: one for lower bounding the $n$, and the other for the private protocol on fixed $n$.

Given a multi-message $(\epsilon, \delta_a)$-DP protocol for a fixed population $n_1$, let $K_{n_1} : \mathcal{X} \mapsto \mathcal{T}^*$ denote its procedures on each client's side. When the population size is $n_2$ ($n_2 \geq n_1$), a natural conclusion according to the post-processing property of DP is that the shuffled output $\bigcup_{i=1}^{n_2} K_{n_1}(\mathbf{x}_i)$ also satisfies $(\epsilon, \delta_a)$-DP. Relying on this observation, we now give a direct approach for multi-message shuffle privacy with a random population in Theorem 5. The proof of the theorem is a simple application of the union bound on violating probabilities.

*Theorem 5.* In the shuffle privacy model where one user participated with certain (denoted as user 1) and the resting population size $n - 1 \sim \mathcal{P}_{N-1}$, if the following conditions are satisfied:

a). there is a shuffle $(\epsilon, \delta_a)$-DP protocol for fixed population size $n_1$ with user's procedure $K_{n_1} : \mathcal{X} \mapsto \mathcal{T}^*$;

b). with probability at least $1 - \delta_b$, the $n - 1 \geq n_1 - 1$,

TABLE II
LIST OF COMMON LEFT LOWER BOUNDS

| **Distribution of** $n-1$ | **With probability** $1-\delta_a$ |
|---|---|
| $Binomial(n', \alpha)$ | $n \geq n'\alpha - \sqrt{n'\alpha \log(1/\delta_a)} + 1$ |
| $Poisson(\gamma)$ | $n \geq \lambda - \sqrt{3\lambda \log(1/\delta_a)} + 1$ |
| $\tau$-sub-Gaussian | $n \geq \mu - \sqrt{\tau^2 \log(1/\delta_a)} + 1$ |
| $(\alpha, \beta)$-sub-exponential | $n \geq \mu - \frac{\log(\alpha/\delta_a)}{\beta} + 1$ |
| bounded $\tau^2$-variance | $n \geq \mu - \sqrt{\frac{\tau^2}{\delta_a}} + 1$ |

then the output $\bigcup_{i=1}^{n} K_{n_1}(\mathbf{x}_i)$ satisfies $(\epsilon, \delta_a + \delta_b)$-DP for the user 1 (w.r.t. the change of user 1's value).

Several efficient multi-message protocols have been proposed in the literature, some of them preserve pure $(\epsilon, 0)$-DP (e.g., in [17]), and others preserve approximate $(\epsilon, \delta_2)$-DP (e.g., in [15], [26], [29]). Given a pure DP protocol, for achieving $(\epsilon, \delta)$-DP with random population, according to the Theorem 5, we can use the whole violating probability $\delta$ for lower bounding $n$. Given an approximate $(\epsilon, \delta_b)$-DP protocol, we can evenly split the violating probability and assign $\delta_a = \delta_b = \frac{\delta}{2}$. For the convenience of referring, here we further list lower $(1-\delta_a)$-bounds of some commonly encountered population distributions in Table II.

*2) Improved Approach Via Expected Bound:* Though Theorem 2 works for any random population distributions with known left lower bounds, its union violating bound $\delta_a + \delta_b$ is usually loose. It simply treats the violating probability of $(\epsilon, 0)$-DP as $\delta_2$ for any $n \geq n_1$. For example, by adopting a secure multiparty shuffling & summation protocol for $n_1$ users with pair-wise networking [11], if each user injects $Gaussian\left(0, \frac{\sqrt{\log(1/\delta)/\epsilon}}{\sqrt{n_1}}\right)$ into their binary value $\mathbf{x}_i \in \{0, 1\}$, then the server observes their summation roughly with $(\epsilon, \delta)$-DP (ignoring negligible incorrectness probability of [11]); when the number of users is $n \geq n_1$, the violating probability of $(\epsilon, 0)$-DP is actually $\delta^{-n/n_1}$, which is less than $\delta$.

Now consider $n - 1 \sim \mathcal{P}_{N-1}$ and fix the user's randomization procedure as $K$, let $(\epsilon, \delta_n)$ denote the DP level of $n$ users' messages $\bigcup_{i=1}^{n} K(\mathbf{x}_i)$, it is easy to observe that the outputting messages of the random population are $(\epsilon, \sum_{n=1}^{+\infty} \mathcal{P}_{N-1}(n) \cdot \delta_n)$-DP as a whole.

We aim to design a multi-message shuffle private protocol for summation with a random participating population (in Algorithms 1 and 2). Here we start with designing a protocol for fixed population. It is based on the general framework for achieving multi-message secure & private summation: split the local value $x_i \in [0.0, 1.0]$ into $m$ additive shares and then shuffle them up [26], [40].

Let $\mathcal{S}$ denote the procedure of shuffling (on $n$ users' shares), and let $\mathcal{R}_{0,m,p,q}$ denote the procedure of Algorithm 1 without adding noises, the seminar work [40] shows that $\mathcal{S} \circ \mathcal{R}_{0,m,p,q}$ with $m \geq 2 + 5\lceil \log(q) \rceil + \lceil \sigma + 2\log(n-1) \rceil$ messages is an $\sigma$-secure protocol on summation (see Definition 6). The recent work [26] further shows $m \geq \left\lceil \frac{2\sigma + \log_2(q)}{\log_2(n) - \log_2(e)} + 1 \right\rceil$ messages that decrease with the number of party $n$ is enough when $m \geq 3$ and $n \geq 19$. Combining these two results on the number of messages, we prove the existence of an efficient $\sigma$-secure protocol with $m = \max\left\{8 + 5\lceil \log(q) \rceil + \lceil \sigma \rceil, 1 + \left\lceil \frac{4\sigma + 2\log_2(q)}{5} \right\rceil\right\}$ messages for any $n$ (in Proposition 1).

**Algorithm 1:** User-Side Randomization.$\mathcal{R}_{\tau,m,p,q}$

```
Input: Local data $x_i \in [0.0, 1.0]$, noise scale $\tau$, number of
  messages $m$ per user, precision $p$, order of the additive
  group $q \geq n' \cdot p$
Output: $y_i^1, \ldots, y_i^m \in [0 : q-1]$
1: ▷ Discretize $x_i$ with precision $p$
2: $\tilde{x}_i = \lfloor x_i \cdot p \rfloor + Bernoulli(x_i \cdot p - \lfloor x_i \cdot p)$
3: ▷ Add discrete Gaussian noise
4: $y_i = \tilde{x}_i + DiscreteGaussian(0, \tau)$
5: ▷ Create random shares
6: Sample $[y_i^1, \ldots, y_i^m] = Uniform([0 : q-1]^m)$
     conditioned on $\sum_{j=1}^{m} y_i^j \equiv y_i$
7: return $\{y_i^1, \ldots, y_i^m\}$
```

*Definition 6 ($\sigma$-secure protocol on summation [40]).* A protocol $\mathcal{M} : \mathrm{Z}_q^n \mapsto \mathrm{O}$ is $\sigma$-secure on summation if for any datasets $D, D^* \in \mathrm{Z}_q^n$ that $\left(\sum_{y \in D} y\right)$ $mod\ q = \left(\sum_{y' \in D^*} y'\right)$ $mod\ q$, the inequality $\frac{1}{2}\sum_{o \in \mathrm{O}} |\mathrm{P}[\mathcal{M}_s(D) = o] - \mathrm{P}[\mathcal{M}_s(D^*) = o]| \leq 2^{-\sigma}$ holds.

*Proposition 1 (Existence of efficient $\sigma$-secure protocols).* There exists an $\sigma$-secure protocol that computes summation over $\mathrm{Z}_q$ in the shuffle model using $m = \max\left\{8 + 5\lceil \log(q) \rceil + \lceil \sigma \rceil, 1 + \left\lceil \frac{4\sigma + 2\log_2(q)}{5} \right\rceil\right\}$ messages for any fixed $n \in \mathcal{Z}_+$ parties.

With the help of an $\sigma$-secure channel on summation, if the summation itself is further sanitized with $(\epsilon, \delta)$-DP (i.e., when $\tau > 0$ in Algorithm 1), then the messages passing through the channel is $(\epsilon, \delta + (1 + e^{\epsilon}) \cdot \sigma)$-DP (see Lemma 3). Therefore, if the summation $\sum_{i=1}^{n} y_i$ from Algorithm 1 is $(\epsilon, \delta)$-DP, then we obtain an $(\epsilon, (1 + e^{\epsilon}) \cdot \sigma)$-DP shuffle private protocol.

*Lemma 3 (Privacy property over $\sigma$-secure protocol [41], [42]).* For any probabilistic map $F : \mathrm{Z}_q^n \mapsto \mathrm{Z}_q^n$ that the summation $\sum F(D) \in \mathrm{Z}_q$ is $(\epsilon, \delta)$-DP for any $D \in \mathrm{Z}_q^n$, if a $n$-party protocol $\mathcal{M}_s : \mathrm{Z}_q^n \mapsto \mathrm{O}$ is $\sigma$-secure on summation, then $\mathcal{M}_s \circ F$ is $(\epsilon, \delta + (1 + e^{\epsilon}) \cdot \sigma)$-DP.

The work of [26] proposes adding Pólya noises on local values to ensemble discrete Laplace noises. However, the Pólya random variables (or discrete Laplace variables) are not closed under summation, making the privacy analyses (e.g., of $\delta_n$) for random population prohibitively hard. Therefore, we adopt the discrete Gaussian noises [43], [44], which are almost closed under summation (see Lemma 4).

*Definition 7 ($DiscreteGaussian(0, \tau)$ distribution [43]).* Let $\mu, \tau \in \mathrm{R}$ with $\tau > 0$. The discrete Gaussian distribution with location $\mu$ and scale $\tau$ is denoted $DiscreteGaussian(\mu, \tau)$. It is a probability distribution supported on the integers and defined

**Algorithm 2:** Server-Side Analytic $\mathcal{A}_{m,p,q}$.

```
Input: Number of messages $k$, messages $\{y_*^j\}_{j=1}^k$, noise
  scale $\tau$, number of messages $m$ per user, precision $p$,
  order of the additive group $q \geq \mathrm{Q}_N[1-\exp(-\tau)] \cdot p$.
Output: $z \in \mathrm{R}$.
▷ Compute number of users
$n = k/m$
▷ Add up all shares
$z = \sum_{j=1}^{k} y_*^j \bmod q$
▷ Correct underflow
if $z > \frac{\mathrm{Q}_N[1-\exp -\tau]\cdot p+q}{2}$ then
    $z = z - q$
end if
return $z/p$
```

by (for $x \in \mathrm{Z}$):

$$\Pr[x] = \frac{e^{-(x-\mu)^2/2\tau^2}}{\sum_{x' \in \mathrm{Z}} e^{-(x'-\mu)^2/2\tau^2}}.$$

*Lemma 4 (Privacy property of summation of discrete Gaussian).* For a dataset $D = [y_1, \ldots, y_n] \in \mathrm{Z}_q^n$, if we define a probabilistic map $F : \mathrm{Z}_q^n \mapsto \mathrm{Z}_q^n$ as adding independent $DiscreteGaussian(0, \tau)$ to each element: $F(D) = \big[y_1 + DiscreteGaussian(0, \tau), \ldots, y_n + DiscreteGaussian(0, \tau)\big]$, then $\left(\epsilon, \exp\left(-0.5\left(\frac{\epsilon - p/(2n\tau^2) - 2.5\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}{p/(\tau\sqrt{n}) + 10\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}\right)^2\right)\right)$-DP.

*Proof.* Due to its almost-closed form [44], the summation of $n$ discrete Gaussian variables satisfy $\frac{1}{2}\beta^2$-concentrated differential privacy [45] for $\tilde{x}_i \in [0 : q-1]$ with $\beta$ as:

$$\min\left\{\sqrt{\frac{p^2}{n\tau^2} + 5\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2\frac{k}{k+1}}}, \frac{p}{\tau\sqrt{n}} + 10\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2\frac{k}{k+1}}\right\}.$$

Since any $\frac{1}{2}\beta^2$-concentrated DP mechanism is $(\epsilon, \delta)$-centralized DP for $\epsilon = \frac{1}{2}\beta^2 + \beta \cdot \sqrt{2\log(1/\delta)}$, we then have the summed discrete Gaussian is centralized $\Big(\frac{p^2}{2n\tau^2} + \frac{5}{2}\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2\frac{k}{k+1}} + \Big(\frac{p}{\tau\sqrt{n}} + 10\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2\frac{k}{k+1}}\Big) \cdot \sqrt{2\log(1/\delta)}, \delta\Big)$-DP. Interchange the variation of $\epsilon$ and $\epsilon$, we then have centralized $\left(\epsilon, \exp\left(-0.5\left(\frac{\epsilon - p^2/(2n\tau^2) - 2.5\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}{p/(\tau\sqrt{n}) + 10\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}\right)^2\right)\right)$-DP. □

Combining the Lemmas 4 and 3, we then have a $(\epsilon, \delta + (1 + e^\epsilon) \cdot \sigma)$-DP shuffle private protocol for fixed $n$.

*Proposition 2 (Existence of Efficient Private Protocols for Fixed Population).* When $p = \lceil\sqrt{n}\rceil, q = \lceil 2n^{3/2}\rceil$ and $m = \max\left\{1 + \left\lceil\frac{4\sigma + 2\log_2(q)}{5}\right\rceil, 8 + 5\lceil\log(q)\rceil + \left\lceil\log(1 + e^\epsilon) + \max\left\{\epsilon\tau - 1/\tau, 1/4\right\}^2\right\rceil\right\}$, the $\mathcal{S} \circ \mathcal{R}_{\tau,k,p,q}$ is a $\left(\epsilon, 2\exp\left(-\frac{1}{2}\left(\frac{\epsilon - 1/\tau^2 + 2.5\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}{1/\tau + 10\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}\right)^2\right)\right)$-DP multi-message shuffle private protocol for $n$ users that sends $O(\log(n\tau\epsilon))$ messages per user and each message is $O(\log n)$ bits, and the mean squared error of $\mathcal{A} \circ \mathcal{S} \circ \mathcal{R}_{\tau,k,p,q}$ is $O(\tau^2)$.

*Proof.* According to Lemma 4, we have the summation is $\left(\epsilon, \exp\left(-0.5\left(\frac{\epsilon - 1/\tau^2 - 2.5\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}{1/\tau + 10\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}\right)^2\right)\right)$-DP. Now set $\sigma = \log(1 + e^\epsilon) + \frac{1}{2}\left(\frac{\epsilon - 1/\tau^2 - 2.5\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}{1/\tau + 10\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}\right)^2$, and $k = \max\left\{1 + \left\lceil\frac{4\sigma + 2\log_2(q)}{5}\right\rceil, 8 + 5\lceil\log(q)\rceil + \left\lceil\log(1 + e^\epsilon) + 1/2\max\left\{\epsilon\tau - 1/\tau, 1/4\right\}\right\rceil\right\} \geq \max\left\{1 + \left\lceil\frac{4\sigma + 2\log_2(q)}{5}\right\rceil, 8 + 5\lceil\log(q)\rceil + \lceil\sigma\rceil\right\}$, we have $(1 + e^\epsilon) \cdot \sigma$ equals $\exp\left(-0.5\left(\frac{\epsilon - 1/\tau^2 - 2.5\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}{1/\tau + 10\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}\right)^2\right)$. According to Proposition 1, the $\mathcal{S} \circ \mathcal{R}_{\tau,k,p,q}$ is thus $\left(\epsilon, 2\exp\left(-0.5\left(\frac{\epsilon - 1/\tau^2 - 2.5\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}{1/\tau + 10\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}\right)^2\right)\right)$-DP.

The number of messages per user is $\lceil\log(m)\rceil = O(log(q + \epsilon^2\tau^2)) = O(\log(n\tau\epsilon))$. Each message is $\lceil\log(q)\rceil = \log n$ bits.

The mean squared error (MSE) comes from rounding, discrete Gaussian noises, and underflow/overflow. The rounding error is bounded by $\frac{n}{p^2}$. The noises are bounded by $\frac{n\tau^2}{p^2}$. Since noises are sub-Gaussian, the underflow/overflow probability is at most $2e^{-\frac{2}{n\tau^2}(q-n\cdot p)^2}$, and the error is bounded by $q^2/p^2$. Consequently, the MSE is: $\frac{n}{p^2} + \frac{n\tau^2}{p^2} + 2q^2/p^2 \cdot e^{-\frac{2}{n\tau^2}(q-n\cdot p)^2} = O(\tau^2)$. □

Reversely, when $(\epsilon, \delta)$ is fixed, according to Theorem 2, we have $\tau = O\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)$, the $\mathcal{A} \circ \mathcal{S} \circ \mathcal{R}_{\tau,k,p,q}$ sends $O(\log n + \log(1/\delta))$ messages, and the mean squared error is $O(\frac{\log(1/\delta)}{\epsilon^2})$. The error matches the Gaussian mechanism for centralized DP.

We proceed to consider the privacy guarantee of the protocol with a random population. When the population is random, the violating probability $\delta$ is an expectation regarding the population size distribution $\mathcal{P}_{N-1}$ as $\delta = \mathbb{E}_{n-1\sim\mathcal{P}_{N-1}}\left[2\exp\left(-0.5\left(\frac{\epsilon - p^2/(n\cdot\tau^2) - 2.5\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}{p/(\tau\cdot\sqrt{n}) + 10\sum_{k=1}^{n-1} e^{-2\pi^2\tau^2 k/k+1}}\right)^2\right)\right]$.

Now consider preserving the utility of the protocol $\mathcal{A} \circ \mathcal{S} \circ \mathcal{R}_{\tau,m,p,q}$. When $n$ is fixed, the $q$ needs to be greater than $n \cdot p$, so that the underflow/overflow probability is negligible. However, when $n$ is random and the potential population size $n'$ is extremely large, simply using $q \geq n' \cdot p$ wastes computation and communication in the protocol. Here we show that setting $q \geq \lceil 2n^* \cdot p\rceil$ is enough for neglecting errors due to underflow/overflow (In Theorem 6), where $2n^*$ is the $\arg\min_{x\geq 1}\int_{n=x}^{+\infty} n^2 \mathrm{d}\mathcal{P}_{N-1}(n-1) \leq \tau^2$.

*Theorem 6.* For extra participating population size $n-1 \sim \mathcal{P}_{N-1}$, let $n^*$ denote $\frac{1}{2}\arg\min_{x\geq 1}\int_{n=x}^{+\infty} n^2 \mathrm{d}\mathcal{P}_{N-1}(n-1) \leq \tau^2$. When $p=\left\lceil\sqrt{\mathbb{E}[N]}\right\rceil, q=\lceil 2n^*\cdot p\rceil$ and $m=\max\left\{1+\left\lceil\frac{4\sigma+2\log_2(q)}{5}\right\rceil, 8+5\lceil\log(q)\rceil+\left\lceil\log(1+e^{\epsilon})+\max\left\{\epsilon\tau-1/\tau, 1/4\right\}^2\right\rceil\right\}$, the $\mathcal{S}\circ\mathcal{R}_{\tau,k,p,q}$ is a $\left(\epsilon, 2\mathbb{E}\left[\exp\left(-\frac{1}{2}\left(\frac{\epsilon-p^2/(n\cdot\tau^2)-2.5\sum_{k=1}^{n-1}e^{-2\pi^2\tau^2 k/k+1}}{p/(\tau\cdot\sqrt{n})+10\sum_{k=1}^{n-1}e^{-2\pi^2\tau^2 k/k+1}}\right)^2\right)\right]\right)$-DP multi-message shuffle private protocol that sends $O\left(\log(n^*\tau\epsilon)\right)$ messages per user and each message is $O(\log\mathbb{E}[N])$ bits, and the mean squared error of $\mathcal{A}\circ\mathcal{S}\circ\mathcal{R}_{\tau,k,p,q}$ is $O(\tau^2)$.

*Proof.* The privacy guarantees and communication costs can be easily inferred from Proposition 2. Recall that the MSE comes from rounding, discrete Gaussian noises, and underflow/overflow. The rounding error is bounded by $\mathbb{E}\left[\frac{n}{p^2}\right]=O(1)$. The noises are bounded by $\mathbb{E}\left[\frac{n\tau^2}{p^2}\right]=O(\tau^2)$.

When $q\geq n\cdot p$, the underflow/overflow probability is at most $2e^{-\frac{2}{n\tau^2}(q-n\cdot p)^2}$, and the error is bounded by $q^2/p^2$. When $q\leq n\cdot p$, the underflow/overflow probability is at most 1.0 and the error is bounded by $n^2$. Therefore, the expected error due to underflow/overflow is bounded by $\int_{n=\lceil q/p\rceil}^{+\infty} n^2\mathrm{d}\mathcal{P}_{N-1}(n-1)+\int_{n=1}^{\lceil q/p\rceil-1} 2q^2/p^2\cdot e^{-\frac{2}{n\tau^2}(q-n\cdot p)^2}\mathrm{d}\mathcal{P}_{N-1}(n-1)=O(\tau^2)$ when $q=\lceil 2n^*\cdot p\rceil$.

Putting all pieces together, we have the MSE is bounded by: $\mathbb{E}\left[\frac{n}{p^2}\right]+\mathbb{E}\left[\frac{n\tau^2}{p^2}\right]+O(\tau^2)=O(\tau^2)$. □

In summary, our protocol in Algorithms 1 and 2 differs from previous approaches [26] in three aspects. First, to derive comprehensible privacy bounds when $n$ is random, we utilize discrete Gaussian noises for DP. Second, to achieve $\sigma$-secure with arbitrary population size $n$, we derive an upper bound of required messages $m$ in Proposition 1. Lastly, for balancing utility and computation/communication costs, we infer a distribution-dependent parameter $q\geq\lceil 2n^*\cdot p\rceil$ for the size of the additive group.

Besides the discrete Gaussian considered here, the Poisson distribution (e.g., in [46]) and Binomial distribution (e.g., in [30]) that are closed under summation can also be easily employed for realizing multi-message shuffle private protocols with a random population.

### B. Dummies in Multi-Message Model

In this part, we consider dummy methods of multi-message shuffle protocols. When $k$ non-adaptive/adaptive dummies are added to the true population, high probability lower bounds on $n$ can still be easily derived based on Table II, and then apply to privacy guarantees via union bounds. Similarly, the population distribution $\mathcal{P}_{N^*-1}$ with dummies can directly plugin into the privacy guarantees via expected bound.

Now consider seeking optimal dummy size $k$ for improving utility, when the privacy guarantee via union bound is used, we can reversely derive the $(\epsilon,\delta_\alpha)$-DP protocol for fixed population size (w.r.t. the new lower bound $n$) with Theorem 5, and find the $k$ that minimizes the protocol's estimation error; when the privacy guarantee via expected bound is utilized, we can derive the new magnitude of local noises (e.g., the Gaussian scale $\tau$) with Theorem 6, and find the $k$ that minimizes the expected MSE.

## VII. Experiments

In this section, we experimentally evaluate the performance of the shuffle privacy model with a random population, and aim to answer the following research questions:

*RQ1:* What is the accuracy gap between the local and shuffle models with a random population size?

*RQ2:* What is the accuracy gap between shuffle models with or without dummies?

*RQ3:* What is the accuracy gap between single-message shuffle and multi-message shuffle?

*RQ4:* What is the accuracy gap between the multi-message shuffle model via union bounds and the multi-message shuffle model via expected bounds?

### A. Settings

Without loss of generality, we consider the aggregation task of distribution estimation on binary data $x_i\in\{0,1\}$, where the binary randomized response is an optimal local private mechanism [38]. We also consider mean estimation over numerical vectors (i.e., gradient data estimation in federated learning) that uses *Laplace* mechansim [1] in the local. The source of potential data $[x_i]_{i=1}^{n'}$ comes from both real world and synthesized datasets:

- UCI Adult Dataset [47]: The Adult dataset contains $n'=48842$ individuals, we use the attribute *gender* (16192 females/32650 males) for aggregation.
- MNIST Dataset: The gradient data comes from training a Logistic regression model with 7850 parameters on MNIST dataset, which contains 60000 hand-written digit images each corresponds to one user. We simulate single-round gradient estimation with random clients check-in (i.e., $N-1$ follows binomial distribution).
- Synthetic Data: To cover more cases in real world, we synthesize datasets with the expected population size ranging from 1000 to 125000, whose binary values distribute near-uniformly [0.55,0.45] or heavily imbalanced as $[0.1, 0.9]$.

Covering most practical cases, the centralized privacy level $\epsilon_c$ is assumed to be ranging from 0.001 to 0.5, and the tail violating probability is fixed to $\delta\equiv 10^{-5}$. Every experiment is simulated by 200 times, and the reported results are the average of these simulations.

### B. Competing Approaches

Our experiments mainly consider population size with Binomial and Poisson random distributions, and the possible approaches for shuffle private aggregation with these random distributions include:

- LOCAL: As the worst-case $N-1=0$ is possible, the local privacy approach must setups $\epsilon_l=\epsilon_c$ (without privacy amplification). Note that the optimal dummy size in the naive shuffle model is 0 (since the binary randomized response has a utility-optimal noise distribution), hence this approach covers the classic shuffle model with dummies [37].
- AMPLIFIED: Utilizing the privacy amplification in Section IV, this approach uses enlarged local privacy budget.
- NON-ADAPTIVE: Use the privacy amplification results and optimal dummy size $k^*$ with non-adaptive dummies in Section V-B.
- ADAPTIVE: Use the privacy amplification results and optimal dummy size $k^*$ with adaptive dummies.
- RandomCheckin: The random check-in protocol [31] allows every user decides participation independently (i.e., $N-1$ follows binomial distribution), then adds dummies to form a fixed-size population and derives corresponding privacy amplification bounds.
- MULTI-UNION: Derive the $(1-\delta/2)$ lower bound $n_1$ of the population size, and then utilize the state-of-the-art multi-message shuffle $(\epsilon, \frac{\delta}{2})$-private protocol for fixed $n_1$ with decentralized two-sided geometric noise [26].
- MULTI-EXPECTED: Use the proposed secure & private protocol with decentralized discrete Gaussian noise (in Section VI-A2), which tightly bounds violating probability via expectation over randomness of population size.

### C. Metrics

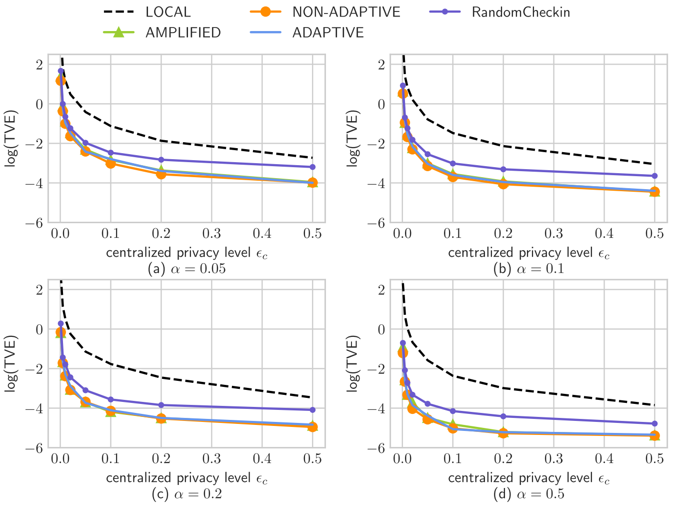
Without projecting the distribution estimator into the probability simplex, since many error metrics are equivalent in binary distribution estimation, here we employ the (natural logarithm of) total variation error (TVE):

$$\mathrm{TVE} = |\mathcal{P}_{\mathcal{X}} - \hat{\mathcal{P}}_{\mathcal{X}}| = \sum_{a\in\mathcal{X}} |\mathcal{P}_{\mathcal{X}}(a) - \hat{\mathcal{P}}_{\mathcal{X}}(a)|.$$

### D. The Effect of Privacy Amplification With Random Population

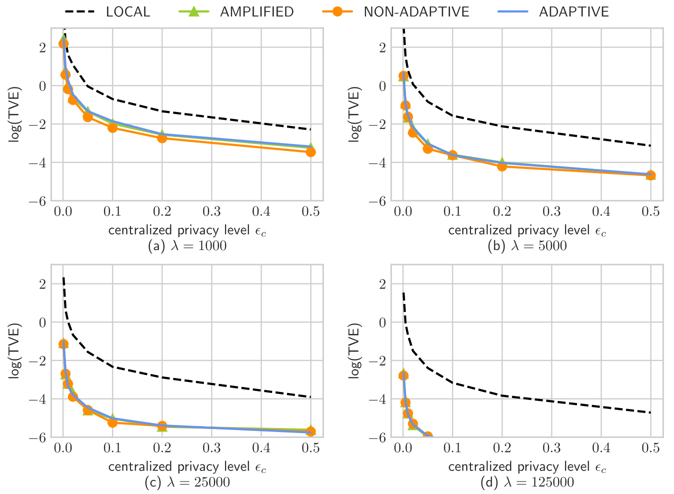
In this part, we study the experimental privacy amplification effects under various random population settings that are commonly encountered in mobile edge computing [48], including the binomial distribution and the Poisson distribution.

With the users in Adult dataset as the potential population, we assume that every user participates with probability $\alpha$ independently. That is, the participating population size $N-1$ follows random distribution $Binomial(48841, \alpha)$. When $\alpha$ is $0.05, 0.1, 0.2,$ and $0.5$, the corresponding experimental results are presented in Fig. 4. It is observed that the AMPLIFIED reduces $70\%$ error compared to the LOCAL approach and the amplification effect increases with the participating rate $\alpha$. The competing RandomCheckin is implemented with $(N-1)\cdot\alpha$ slots (i.e., the number of effective participating messages) as suggested by the original work [31]. As it uniformly selects one user for each slot that multiple users checked-in, and adds dummies to slots that no user checked-in, the RandomCheckin often gives slightly more privacy amplification effect than bounds in Theorems 1 and 8. On the other hand, there are more dummies and fewer non-dummy messages in RandomCheckin, thus the final utility has an average $50\%$ gap to our approaches.



(a) $\alpha=0.05$ (b) $\alpha=0.1$ (c) $\alpha=0.2$ (d) $\alpha=0.5$

Fig. 4. Total variation error results on the Adult dataset with $n'=48841$ and binomial participate rate $\alpha$ ranges from 0.01 to 0.5.



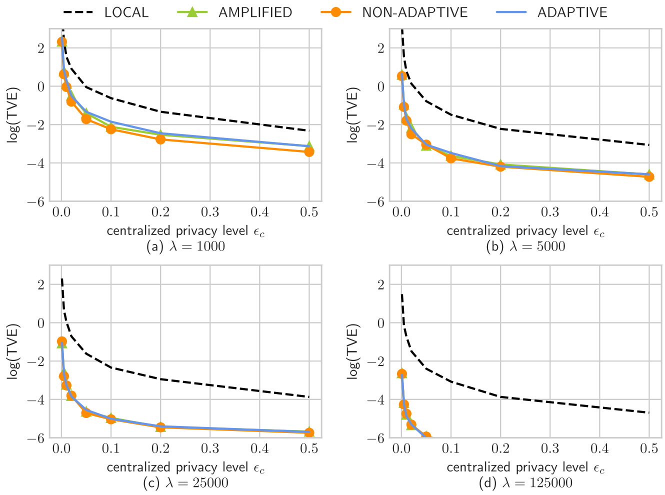
(a) $\lambda=1000$ (b) $\lambda=5000$ (c) $\lambda=25000$ (d) $\lambda=125000$

Fig. 5. Total variation error results on the synthetic dataset with true distribution $[0.55, 0.45]$ and Poisson mean $\alpha$ ranges from 1000 to 125000.

In synthesized experiments, we assume the population size $n'-1$ follows a Poisson distribution. With the mean value $\lambda$ defined in $\{1000, 5000, 250000, 125000\}$, we present experimental results in Fig. 5 (with true data distribution $[0.55, 0.45]$) and 6 (with true data distribution $[0.1, 0.9]$). Generally, the TVE errors of all approaches are linearly with $\frac{1}{\lambda}$. The AMPLIFIED reduces about $60\%$ error compared to the LOCAL approach, and the error reduction gap grows with the expected population size $\lambda$.

### E. The Effect of Dummies

In this part, we study the effect of adding non-adaptive/adaptive dummies. In the experiments on the Adult



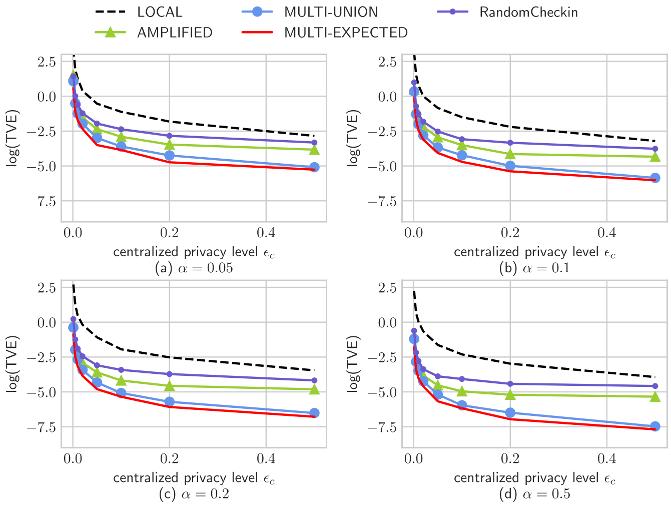
(a) $\lambda = 1000$

(b) $\lambda = 5000$

(c) $\lambda = 25000$

(d) $\lambda = 125000$

Fig. 6. Total variation error results on the synthetic dataset with true distribution $[0.1, 0.9]$ and Poisson mean $\lambda$ ranges from 1000 to 125000.



(a) $\alpha = 0.05$

(b) $\alpha = 0.1$

(c) $\alpha = 0.2$

(d) $\alpha = 0.5$

Fig. 7. Total variation error results on the Adult dataset with $n' = 48841$ and binomial participating rate $\alpha$ ranges from 0.01 to 0.5.

dataset with binomial distribution (i.e., Fig. 4), we observe that the NON-ADAPTIVE approach could further reduce about 10% error compared to AMPLIFIED, when the expected population size and the centralized privacy budget is small (e.g., when $\alpha = 0.05$ and $\epsilon_c = 0.001$). The ADAPTIVE is always dominated by the NON-ADAPTIVE, and has almost the same performances as the AMPLIFIED approach in most cases.

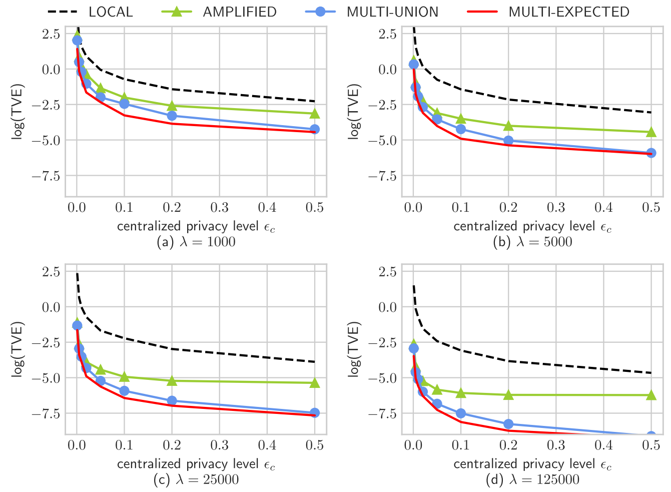
In synthesized experiments (i.e., Figs. 5 and 6), we assume the population size $n' - 1$ follows a Poisson distribution. With the mean value $\lambda$ defined in $\{1000, 5000, 250000, 125000\}$, we present experimental results in Fig. 5 (with true data distribution $[0.55, 0.45]$) and 6 (with true data distribution $[0.1, 0.9]$). When $\lambda$ is relatively small, the NON-ADAPTIVE approach reduces about 10% error compared to the AMPLIFIED. In summary, adding dummies is effective when the expected population size is relatively small.

### F. The Effect of Multi-Message Shuffle Model

Following the same simulation settings, we present experimental results of multi-message shuffle private protocols for the Adult/synthetic datasets in Figs. 7 and 8 respectively. Compared to protocols restricted to transmitting one message, multi-message protocols MULTI-UNION/MULTI-EXPECTED averagely reduce about 60% estimation error. Specifically, the MULTI-EXPECTED outperforms the MULTI-UNION by about 40% in the intermediate privacy regime (e.g., when $0.05 < \epsilon_c < 0.3$). When the privacy budget gets larger, since the two-sided geometric noises introduce less noise than Laplace or Gaussian noises [49], the performance gap between MULTI-EXPECTED and MULTI-UNION narrows down.

### G. Results on Gradient Estimation

To illustrate the effectiveness of our approaches for diverse user data and aggregation tasks, this part dedicates to gradient estimation with binomial clients population (i.e., random check-in) on the MNIST dataset. The gradient vector from each user is of length 7850 and is clipped by $C = 5$ under $\ell_1$-norm, then noised by $Laplace(2\,C/\epsilon_l)$. We present TVE results of single-round gradient vector estimation in Fig. 9 with vary participating rates. Similar to the performance gaps for binary distribution estimation, it is observed that our approaches outperformed the RandomCheckin by about 40%.



(a) $\lambda = 1000$

(b) $\lambda = 5000$

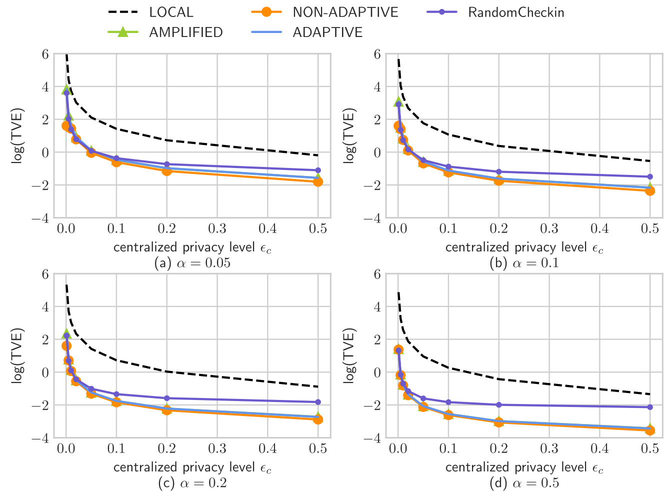
(c) $\lambda = 25000$

(d) $\lambda = 125000$

Fig. 8. Total variation error results on the synthetic dataset with true distribution $[0.55, 0.45]$ and Poisson mean $\alpha$ ranges from 1000 to 125000.

### H. Summary

In the single-message shuffle model, with the privacy amplification bounds derived for a random population, we achieve significant accuracy boosts. Besides, with the optimal adaptive-dummy size, the accuracy can further be slightly improved, which validates our theoretical/numerical analyses. It is observed that the multi-message shuffle private protocols achieve



(a) $\alpha = 0.05$

(b) $\alpha = 0.1$

(c) $\alpha = 0.2$

(d) $\alpha = 0.5$

Fig. 9. Total variation error results of gradient estimation on MNIST dataset, the participating rate $\alpha$ ranges from 0.01 to 0.5.

higher accuracy than single-message ones, though incurs more communication costs (e.g., tens of messages per user).

## VIII. Conclusion

This work initialized the study of shuffle differential privacy with a random participating population, and provided solutions for efficient and accurate private data analyses in highly dynamic networking environments (e.g., mobile computing). For population size with Binomial, Poisson, sub-Gaussian, and general distributions, we have derived privacy amplification lower bounds in the single-message shuffle model. We further give amplification bounds when non-adaptive or adaptive dummies are added, which works as a remedy to the randomness (high tail probabilities) of the population size. Based on these bounds, we then studied optimal dummy sizes to improve the estimation accuracy. Finally, relying on the property of almost closed form of summations on discrete Gaussian noises, we propose a multi-message shuffle private protocol for random population. Through experiments on real-world and synthetic datasets, we show more than 60% accuracy boosts over naive approaches.

*Discussion:* When the randomness of the current user's participating choice is considered, the privacy guarantee can furthered be enhanced by sub-sampling [36]. It is also straight-forward to extend our results to batch/sequential data analyses (e.g., in federated learning) by replacing the binary $0/1$ participating model to a categorical one.

## References

[1] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput.*, Springer, 2008, pp. 1–19.

[2] J. Zhang, X. Xiao, Y. Yang, Z. Zhang, and M. Winslett, "PrivGene: Differentially private model fitting using genetic algorithms," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 665–676.

[3] M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.

[4] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. IEEE 54th Annu. Symp. Found. Comput. Sci.*, 2013, pp. 429–438.

[5] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2014, pp. 1054–1067.

[6] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting telemetry data privately," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3574–3583.

[7] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang, "Privacy loss in apple's implementation of differential privacy on macos 10.12," 2017, *arXiv:1709.02753*.

[8] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by shuffling: From local to central differential privacy via anonymity," in *Proc. 13th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2019, pp. 2468–2479.

[9] A. Bittau et al., "Prochlo: Strong privacy for analytics in the crowd," in *Proc. 26th Symp. Operating Syst. Princ.*, 2017, pp. 441–459.

[10] M. Abe, "Universally verifiable mix-net with verification work independent of the number of mix-servers," in *Proc. Int. Conf. Theory Appl. Cryptographic Techn.*, Springer, 1998, pp. 437–447.

[11] J. H. Bell, K. A. Bonawitz, A. Gascón, T. Lepoint, and M. Raykova, "Secure single-server aggregation with (poly) logarithmic overhead," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2020, pp. 1253–1269.

[12] A. Cheu, A. Smith, J. Ullman, D. Zeber, and M. Zhilyaev, "Distributed differential privacy via shuffling," in *Proc. Annu. Int. Conf. Theory Appl. Cryptographic Techn.*, Springer, 2019, pp. 375–403.

[13] B. Balle, J. Bell, A. Gascón, and K. Nissim, "The privacy blanket of the shuffle model," in *Proc. Annu. Int. Cryptol. Conf.*, Springer, 2019, pp. 638–667.

[14] V. Feldman, A. McMillan, and K. Talwar, "Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling," in *Proc. 62nd Annu. Symp. Found. Comput. Sci.*, 2022, pp. 954–964.

[15] A. Cheu and J. Ullman, "The limits of pan privacy and shuffle privacy for learning and estimation," in *Proc. 53rd Annu. ACM SIGACT Symp. Theory Comput.*, 2021, pp. 1081–1094.

[16] B. Ghazi, N. Golowich, R. Kumar, P. Manurangsi, R. Pagh, and A. Velingker, "Pure differentially private summation from anonymous messages," 2020, *arXiv:2002.01919*.

[17] B. Ghazi, R. Kumar, P. Manurangsi, and R. Pagh, "Private counting from anonymous messages: Near-optimal accuracy with vanishing communication overhead," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 3505–3514.

[18] R. Liu, Y. Cao, H. Chen, R. Guo, and M. Yoshikawa, "Flame: Differentially private federated learning in the shuffle model," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 8688–8696.

[19] A. Girgis, D. Data, S. Diggavi, P. Kairouz, and A. T. Suresh, "Shuffled model of differential privacy in federated learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 2521–2529.

[20] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. IEEE INFOCOM*, 2011, pp. 882–890.

[21] R. Stanica, M. Fiore, and F. Malandrino, "Offloading floating car data," in *Proc. IEEE 14th Int. Symp. A World Wireless Mobile Multimedia Netw.*, 2013, pp. 1–9.

[22] B. Ghazi, N. Golowich, R. Kumar, R. Pagh, and A. Velingker, "On the power of multiple anonymous messages," 2019, *arXiv:1908.11358*.

[23] B. Liu, Y. Li, Y. Liu, Y. Guo, and X. Chen, "PMC: A privacy-preserving deep learning model customization framework for edge computing," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 4, no. 4, pp. 1–25, 2020.

[24] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 2436–2444.

[25] S. Wang et al., "Local differential private data aggregation for discrete distribution estimation," *IEEE Trans. Parallel Distrib. Syst.*, vol. 30, no. 9, pp. 2046–2059, Sep. 2019.

[26] B. Balle, J. Bell, A. Gascón, and K. Nissim, "Private summation in the multi-message shuffle model," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2020, pp. 657–676.

[27] A. Cheu and J. Ullman, "The limits of pan privacy and shuffle privacy for learning and estimation," 2020, *arXiv:2009.08000*.

[28] A. Koskela, M. A. Heikkilä, and A. Honkela, "Tight accounting in the shuffle model of differential privacy," 2021, *arXiv:2106.00477*.

[29] B. Ghazi, R. Kumar, P. Manurangsi, R. Pagh, and A. Sinha, "Differentially private aggregation in the shuffle model: Almost central accuracy in almost a single message," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 3692–3701.

[30] X. Li et al., "Dump: A dummy-point-based framework for histogram estimation in shuffle model," 2020, *arXiv:12009.13738*.

[31] B. Balle, P. Kairouz, B. McMahan, O. D. Thakkar, and A. Thakurta, "Privacy amplification via random check-ins," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 4623–4634.

[32] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Proc. Int. Conf. Theory Appl. Cryptographic Techn.*, Springer, 1999, pp. 223–238.

[33] T. Wang et al., "Improving utility and security of the shuffler-based differential privacy," *Proc. VLDB Endowment*, vol. 13, pp. 3545–3558, 2020.

[34] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. Amer. Statist. Assoc.*, vol. 60, no. 309, pp. 63–69, 1965.

[35] C. Dwork et al., "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3/4, pp. 211–407, 2014.

[36] B. Balle, G. Barthe, and M. Gaboardi, "Privacy profiles and amplification by subsampling," *J. Privacy Confidentiality*, vol. 10, no. 1, 2020.

[37] T. Wang et al., "Improving utility and security of the shuffler-based differential privacy," 2019, *arXiv:1908.11515*.

[38] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2879–2887.

[39] S. Goryczka and L. Xiong, "A comprehensive comparison of multiparty secure additions with differential privacy," *IEEE Trans. Dependable Secure Comput.*, vol. 14, no. 5, pp. 463–477, Sep./Oct. 2017.

[40] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, "Cryptography from anonymity," in *Proc. 47th Annu. IEEE Symp. Found. Comput. Sci.*, 2006, pp. 239–248.

[41] Y.-X. Wang, S. Fienberg, and A. Smola, "Privacy for free: Posterior sampling and stochastic gradient monte carlo," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2493–2502.

[42] B. Balle, J. Bell, A. Gascon, and K. Nissim, "Differentially private summation with multi-message shuffling," 2019, *arXiv:1906.09116*.

[43] C. L. Canonne, G. Kamath, and T. Steinke, "The discrete gaussian for differential privacy," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 15676–15688.

[44] P. Kairouz, Z. Liu, and T. Steinke, "The distributed discrete gaussian mechanism for federated learning with secure aggregation," 2021, *arXiv:2102.06387*.

[45] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," 2016, *arXiv:1603.01887*.

[46] N. Agarwal, P. Kairouz, and Z. Liu, "The skellam mechanism for differentially private federated learning," 2021, *arXiv:2110.04995*.

[47] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid," in *Knowl. Discov. Data Mining*, vol. 96, pp. 202–207, 1996.

[48] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Commun. Surv. Tut.*, vol. 15, no. 3, pp. 996–1019, Third Quarter 2013.

[49] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally utility-maximizing privacy mechanisms," *SIAM J. Comput.*, vol. 41, no. 6, pp. 1673–1693, 2012.

[50] J. V. Terza and P. W. Wilson, "Analyzing frequencies of several types of events: A mixed multinomial-poisson approach," *Rev. Econ. Statist.*, vol. 72, pp. 108–115, 1990.

[51] A. R. Zhang and Y. Zhou, "On the non-asymptotic and sharp lower tail bounds of random variables," *Stat*, vol. 9, no. 1, 2020, Art. no. e314.

[52] G. Bennett, "Probability inequalities for the sum of independent random variables," *J. Amer. Statist. Assoc.*, vol. 57, no. 297, pp. 33–45, 1962.

**Shaowei Wang** received the PhD degree in the School of Computer Science and Technology from the University of Science and Technology of China (USTC), in 2019. He is an associate professor in Institute of Artificial Intelligence and Blockchain with Guangzhou University. His research interests are data privacy, federated learning and recommendation systems. He has published more than 20 papers on top-tier conferences and journals, such as INFOCOM, VLDB, IJCAI, *IEEE Transactions on Parallel and Distributed Systems*, and *IEEE Transactions on Knowledge and Data Engineering*.

**Xuandi Luo** is currently working toward the graduation degree with Guangzhou University. His research interests are data privacy and federated learning.

**Yuqiu Qian** received the PhD degree in computer science from the University of Hong Kong, in 2019. She was an applied researcher in Tencent, Shenzhen, China. She is currently a solutions architect in Amazon Web Services, Shenzhen, China.

**Youwen Zhu** received the PhD degree in computer science from the University of Science and Technology of China, in 2012. He is currently a professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. His research interests include identity authentication, information security and data privacy.

**Kongyang Chen** received the PhD degree in computer science from the University of Chinese Academy of Sciences, China. He is currently an associate professor with the Institutes of Artificial Intelligence and Blockchain, Guangzhou University, China. His main research interests are artificial intelligence, privacy computing, edge computing, as well as distributed systems such as Internet of Things (IoT) and blockchain.

**Qi Chen** received the PhD degree in mathematics from the Guangzhou university, China, in 2011. Since 2017, he has been with Guangzhou University. His research interests include secret sharing, cryptography, and coding theory.

**Bangzhou Xin** is currently working toward the PhD degree in the School of CyberScience, University of Science and Technology of China. His research interests include data privacy and machine learning.

**Wei Yang** (Member, IEEE) received the PhD degree in computer science from the University of Science and Technology of China (USTC), in 2007. He is an associate professor with the School of Computer Science and Technology, USTC. His research interests include information security, quantum information, and human-computer interaction.