High-Accuracy, Poisoning-Resilient Frequency Estimation in the Shuffle Model

Shaoqiang Wu[†*], Jingyu Jia[†‡*], Yikuan Zhu[†], Xinhao Li[†], Changyu Dong[§], Zheli Liu[†✉]

[†] *CS&CCS, DISSec, AAIS, Nankai University, Tianjin, China,*

{wushaoqiang, jiajingyu, yikuanzhu, asunalxh}@mail.nankai.edu.cn, liuzheli@nankai.edu.cn

[‡] *Automotive New Technology Research Institute of BYD, Shenzhen, China*

[§] *Guangzhou University, Guangzhou, China, Changyu.dong@gmail.com*

Abstract

We study frequency estimation in the shuffle model of differential privacy under *poisoning attacks*, where corrupted users may deviate from the local randomizer to inject crafted in-domain messages. Existing shuffle-model protocols face a core tension: achieving low estimation error relies on flexible multi-message noise generation, which can amplify poisoning influence once messages are anonymized by shuffling.

To address this tension, we propose a *symmetric binomial-sum noise distribution* (i.e., $\mathrm{Bin}(n/2,p)+\mathrm{Bin}(n/2,1-p)$), which preserves high accuracy while limiting the impact of crafted in-domain messages. We realize this distribution via preprocessing-guided noise generation, which routes a balanced collection of mode flags through the shuffler so that each user receives a randomly assigned mode flag that fixes their noise-sampling behavior prior to shuffling. For binary estimation, our protocol requires a single Bernoulli trial per user and at most 2 messages per user (1.5 on average), while bounding the worst-case poisoning influence of a single corrupted user by $O(1/n)$. We extend the protocol to histograms, including large domains via hashing, and provide formal privacy, accuracy, and robustness guarantees. Experiments on real datasets show that our protocols remain resilient under poisoning and reduce MAE by up to nearly $2\times$ over the strongest baseline at comparable per-user communication on small domains, and match it on large domains.

1 Introduction

Differential privacy (DP) has seen broad real-world deployment by both governments and industry [1, 15, 20]. The shuffle model of differential privacy improves the trust–accuracy trade-off by combining local randomization with *privacy amplification by shuffling*, which anonymizes reports by breaking the linkage between senders and messages [4, 5, 10]. For basic analytics tasks such as frequency estimation and histogram

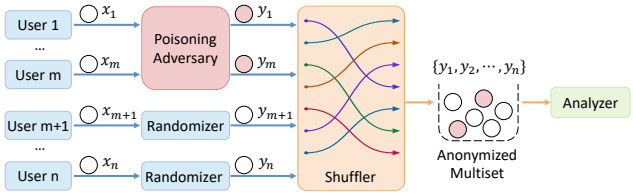

Figure 1: Poisoning attacks in the shuffle model.

construction, a growing body of work develops shuffle protocols that achieve near-central accuracy under similar privacy parameters [2, 11, 18, 19, 23].

After shuffling, the analyzer observes only an unlabeled multiset and, in the standard model, cannot attribute messages to users, a loss that is particularly consequential when some users are adversarial and deviate from the prescribed local randomizer (Fig. 1). While DP bounds how much the output distribution of a fixed mechanism can change under a single-record change, it does not, by itself, imply robustness to arbitrary malicious deviations from the prescribed local randomizer. Many accuracy-oriented shuffle protocols rely on *multi-message* reports to realize low-variance aggregate noise; under anonymity, this multi-message freedom can be exploited for poisoning. Even with a per-user message cap, if many syntactically valid multisets fit within the cap, corrupted users can choose the multiset that maximizes the shift in the estimator's expectation. As a result, existing designs face a tension. Tightly constrained designs admit explicit worst-case poisoning bounds at the cost of higher error, whereas accuracy-oriented multi-message designs achieve lower error but do not provide such bounds in the standard shuffle model [11, 23].

We study shuffle-model frequency estimation under *poisoning attacks* in which an adversary corrupts up to m users and makes them deviate arbitrarily, subject only to the protocol's *per-user message limit*. We quantify robustness by *poisoning influence*: the maximum ℓ_1 shift in the *expected* estimate induced by m corrupted users. Our goal is to design

[*] Equal contribution. ✉ Corresponding author.

shuffle protocols that simultaneously (i) satisfy (ε, δ)-DP, (ii) improve on the accuracy of state-of-the-art multi-message shuffle protocols at comparable per-user communication, and (iii) provide an explicit upper bound on poisoning influence, without assuming shuffler capabilities beyond permutation.

Our main technical contribution is a new aggregate-noise design for the shuffle model: the *symmetric binomial-sum noise distribution* $\mathrm{Bin}(n/2, p) + \mathrm{Bin}(n/2, 1-p)$, together with a shuffle-only generation method that is resilient to poisoning. At a high level, we introduce a *data-independent* preprocessing step that assigns each user a mode flag $b_i \in \{0,1\}$ hidden from the analyzer. Concretely, the analyzer submits to the shuffler a multiset containing $n/2$ copies of 0 and $n/2$ copies of 1; the shuffler uniformly permutes these mode flags, and each user receives one permuted mode flag as auxiliary input to the local randomizer, while the analyzer never observes the per-user assignment. Conditioned on b_i, each user samples noise using one of two complementary probabilities (p vs. $1-p$), so that the *aggregate* noise follows the symmetric binomial-sum form. Crucially, the same mechanism that shapes aggregate noise also shrinks the adversary's action space under anonymity, restricting each user's in-domain multiset to a small family (e.g., $\{0,1,2\}$ copies in the binary protocol), which bounds worst-case bias.

This design yields two concrete benefits. First, compared to binomial or skewed binomial-variant noise commonly induced by bounded-influence shuffle protocols [11, 23], our analysis and experiments show that the resulting symmetric binomial-sum noise achieves improved accuracy under comparable privacy and communication. Second, because preprocessing limits each user's feasible in-domain multiset, each additional in-domain message has a bounded marginal effect on the estimate; moreover, our estimators apply a constant shift rather than an amplifying multiplicative debiasing, so adversarial perturbations are not magnified. We provide a complete theoretical analysis of this symmetric binomial-sum noise design, establishing (ε, δ)-DP and explicit accuracy guarantees for binary estimation and for histograms, and deriving explicit poisoning-influence bounds under the standard per-user message-limit premise.

We instantiate the design in a family of frequency-estimation protocols. In the binary setting, each user can contribute only a small in-domain multiset (at most two copies of the symbol 1), and the analyzer estimates the mean via the constant-shift estimator $\tilde{f} = |\mathbf{Y}|/n - 1/2$, where $|\mathbf{Y}|$ is the total number of shuffled messages. The protocol uses at most 2 messages per user (and 1.5 on average). We then extend the design to histograms. A per-bin instantiation incurs $O(d)$ messages per user, so we develop a low-communication protocol in which each user is assigned a single (hidden) noise bin (and a mode) and injects noise only into that bin. This removes the linear-in-d communication cost, and each message carries a $\log d$-bit bin label, while each user perturbs only one bin. Finally, to cover large domains, we give a compressed variant that hashes values into a smaller domain, performs aggregation in the hashed space, and reconstructs unbiased estimates with bounded collision error. Tab. 1 compares our protocols with representative prior work in terms of communication, accuracy, and poisoning influence.

Contributions. We make the following contributions:

- **A new symmetric binomial-sum noise construction.** We introduce a symmetric binomial-sum noise distribution and show how to realize it with a permutation-only shuffler via a data-independent mode assignment, enabling symmetric, low-variance aggregate noise while reducing the poisoning adversary's degrees of freedom under anonymity.
- **Poisoning-resilient protocols with improved accuracy.** Building on the new noise construction, we design binary and histogram protocols that use constant-shift estimators and yield explicit poisoning-influence bounds under the standard per-user message-limit premise, while improving accuracy over binomial or skewed binomial-variant bounded-influence designs.
- **Theory and empirics.** We provide formal analyses of privacy, accuracy, poisoning influence, and message complexity, and empirically validate the resulting accuracy–communication–poisoning-robustness trade-offs on real datasets. Experiments indicate that we reduce MAE by up to nearly $2\times$ over the best prior protocol [23] at comparable per-user communication and poisoning robustness on the small-domain datasets, and match it on the large-domain dataset.

2 Related Work

Shuffle-model frequency estimation. Frequency estimation and histogram construction have been widely studied in the shuffle model. A first line of work studies *single-message* protocols, where each user locally randomizes their input and sends one report to the shuffler. Shuffling amplifies privacy and can reduce the noise needed relative to purely local mechanisms [4, 10], yet single-message protocols still exhibit a provable accuracy gap to the central model [10].

A second line of work improves accuracy by allowing each user to send *multiple* anonymous messages [2, 5, 17, 18]. To approach central-model error, Balle et al. [5] realize low-variance (e.g., discrete-Laplace-like) noise via split-and-mix messages [21]. Later refinements improve communication by altering the noise-generation/encoding procedure, typically trading a small accuracy loss for fewer messages [18, 19]. However, once messages are anonymized, a corrupted user can choose an arbitrary multiset of in-domain messages within the per-user message limit, which can translate into a large

Table 1: Comparison of shuffle-model frequency-estimation protocols under poisoning attacks. We use (ε, δ) for privacy, n for the number of users, d for the histogram domain size, and d_h for the hashed domain size. *Message per user* is the average number of messages sent by an honest user. *Max error* is a constant-success-probability bound (up to constants) on the ℓ_∞ deviation of the frequency statistic. *Poisoning influence* is the maximum ℓ_1 deviation in the expected frequency statistic under m corrupted users subject to the per-user message limit.

Task	Protocol	Message per user	Max error	Poisoning influence
Binary	CSUZZ [10]	1	$O\left(\frac{\sqrt{\log(1/\delta)}}{\varepsilon n}\right)$	$O\left(\frac{m}{n}/(1-\frac{\sqrt{\log(1/\delta)}}{\varepsilon^2 n})\right)$
Binary	BC [2]	$2-\frac{50\log(2/\delta)}{\varepsilon^2 n}$	$O\left(\frac{\log(1/\delta)}{\varepsilon^2 n}\right)$	$O\left(\frac{m}{n}+\frac{\log(1/\delta)}{\varepsilon^2 n}\right)$
Binary	BBGN [5]	$O(\log\log n)$	$O\left(\frac{\log\log n\sqrt{\log(1/\delta)}}{\varepsilon n}\right)$	$O\left(\frac{m}{n}\cdot(1+\frac{(\log\log n)^2\log(1/\delta)}{n^{2/3}\varepsilon^2})\right)$
Binary	Ours (Thm. 3)	1.5	$O\left(\frac{\sqrt{\log(1/\delta)}}{\varepsilon n}\right)$	$O\left(\frac{m}{n}\right)$
Histogram	CSUZZ [10]	d	$O\left(\frac{\sqrt{\log d\cdot\log(1/\delta)}}{\varepsilon n}\right)$	$O\left(\frac{m}{n}\cdot d/(1-\frac{\sqrt{\log(1/\delta)}}{\varepsilon^2 n})\right)$
Histogram	BC [2]	$1+d-\frac{200d\log(4/\delta)}{\varepsilon^2 n}$	$O\left(\frac{\log(1/\delta)}{\varepsilon^2 n}\right)$	$O\left(\frac{m}{n}\cdot(1+d)\right)$
Histogram	GKMP [18]	$1+O\left(\frac{d\log^2(1/\delta)}{\varepsilon^2 n}\right)$	$O\left(\frac{\log d}{\varepsilon n}\right)$	unbounded†
Histogram	GGKPV [17]	$O\left(\frac{\log^3 d\ \log(\log d/\delta)}{\varepsilon^2}\right)$	$O\left(\frac{\log^{3/2} d\sqrt{\log(\log d/\delta)}}{\varepsilon n}\right)$	$O\left(\frac{m}{n}\cdot\frac{\log^2 d\ \log(\log d/\delta)}{\varepsilon^2}\right)$
Histogram	CZ [11]	$1+O\left(\frac{\log(1/\delta)}{\varepsilon^2 n}\right)$	$O\left(\frac{\log d}{n}+\frac{\sqrt{\log d\cdot\log(1/\delta)}}{\varepsilon n}\right)$	$O\left(\frac{m}{n}\cdot d\cdot(1+\frac{\log(1/\delta)}{\varepsilon^2 n}+\frac{\log d}{n})\right)$
Histogram	LWY (small d) [23]	$1+\frac{32d\log(2/\delta)}{\varepsilon^2 n}$	$O\left(\frac{\log d}{n}+\frac{\sqrt{\log d\cdot\log(1/\delta)}}{\varepsilon n}\right)$	$O\left(\frac{m}{n}\cdot(1+\frac{d\log(1/\delta)}{\varepsilon^2 n})\right)$
Histogram	LWY (large d) [23]	$1+\frac{32d_h\log(2/\delta)}{\varepsilon^2 n}$	$O\left(\frac{\log d}{n}+\frac{\sqrt{\log d(\varepsilon^2 n/d_h+\log(1/\delta))}}{\varepsilon n}\right)$	$O\left(\frac{m}{n}\cdot d\cdot(1+\frac{d_h\log(1/\delta)}{\varepsilon^2 n})\right)$
Histogram	Ours (Thm. 7, small d)	$1+O\left(\frac{d\log(1/\delta)}{\varepsilon^2 n}\right)$	$O\left(\frac{\sqrt{\log d\cdot\log(1/\delta)}}{\varepsilon n}\right)$	$O\left(\frac{m}{n}\cdot(1+\frac{d\log(1/\delta)}{\varepsilon^2 n})\right)$
Histogram	Ours (Thm. 8, large d)	$1+O\left(\frac{d_h\log(1/\delta)}{\varepsilon^2 n}\right)$	$O\left(\sqrt{\frac{\log d}{nd_h}}+\frac{\sqrt{\log d\cdot\log(1/\delta)}}{\varepsilon n}\right)$	$O\left(\frac{m}{n}\cdot d\cdot(1+\frac{d_h\log(1/\delta)}{\varepsilon^2 n})\right)$

† GKMP has no deterministic per-user message limit. Its noise count is unbounded, so any external limit distorts honest behavior, while no limit leaves poisoning unbounded. GKMP's message-per-user bound above hides a large constant under its accuracy-favoring privacy-budget split.

worst-case bias of the final estimate. In contrast, another family of multi-message protocols enforces a per-user message limit and a restricted message domain, typically inducing binomial (or binomial-variant) noise. Such restrictions make it possible to derive an explicit per-user influence bound, but may increase estimation error when the induced noise distribution is asymmetric or otherwise mismatched to the target symmetric noise [11, 23].

Closest baselines under the standard shuffle model. Luo et al. [23] provide our closest baseline: a blanket-noise design in which each user sends the true item and, with a data-independent Bernoulli probability, additionally sends a uniformly sampled domain element. Our design instead uses data-independent preprocessing to assign each user a specific noise bin and mode flag, inducing a symmetric binomial-sum noise distribution at each bin. This comparison therefore highlights the effect of symmetric noise generation on accuracy, and the narrower tails of our symmetric binomial-sum noise help explain the lower MAE observed in Tab. 2.

Poisoning and robustness. Poisoning/manipulation attacks have been systematically studied in the local model, showing that adversarial users can exploit local randomizers to submit crafted yet *in-domain* reports that bias the final estimate [7, 9]. Subsequent work extends these attacks to richer tasks (e.g., key–value, mean and variance estimation) and further shows that the effect of privacy strength on manipulation can be

subtle and mechanism-dependent [22,28]. These concerns are also relevant to shuffle protocols built from local randomizers, since by removing sender–message linkage, shuffling can make it harder to enforce or audit per-user reporting structure and to detect structured manipulations once reports appear syntactically valid.

Recent work by Murakami et al. [24] addresses poisoning by moving beyond the standard shuffle setting. They study an *augmented* shuffle model where the shuffler performs additional online operations beyond permutation, namely random sampling and dummy-data addition over user-encrypted inputs. These added capabilities enable local-noise-free constructions and robustness guarantees against poisoning/collusion, but they rely on a stronger shuffler model and incur cryptographic and communication overhead; moreover, the overhead can become challenging to scale to large domains. In contrast, our protocol uses only the shuffler's standard permutation operation, including the components that realize the model extension. The model extension we adopt (Def. 3) is confined to allowing each user to receive a data-independent auxiliary input before the local randomizer applies. Our robustness, in turn, comes from structural constraints on user reports rather than added shuffler functionality, yielding explicit poisoning-influence bounds.

A related but orthogonal line studies *robust shuffle privacy* [3], which concerns privacy robustness. It characterizes how privacy guarantees degrade when only a fraction of users execute the prescribed protocol while the rest may behave arbitrarily. For example, Luo et al. [23] discuss instantiations that tolerate a constant fraction of such users under this privacy notion. In contrast, we focus on bounding the estimator's worst-case bias under adversarial in-domain reports, given a per-user message limit.

Our work occupies a distinct point in the shuffle-protocol design space. First, unlike augmented-shuffle approaches [24], our shuffler is permutation-only (no online sampling or dummy injection) and we derive explicit poisoning-influence bounds under this weaker shuffler capability; we target bounded estimator bias under a per-user message limit, rather than the privacy-degradation framing of robust shuffle privacy. Second, compared to many accuracy-oriented multi-message shuffle protocols [2,5,17,18], where anonymity can leave corrupted users substantial freedom to choose protocol-conforming in-domain message multisets within the message cap, we co-design the noise and report syntax to achieve *both* high accuracy and bounded poisoning influence. Third, within bounded-influence multi-message designs based on binomial or skewed binomial-variant noise [11,23], our symmetric binomial-sum noise avoids the accuracy penalty of such asymmetric noise. In sum, both high accuracy and poisoning resilience follow from this design, not from a stronger shuffler or a looser per-user message limit.

3 Preliminaries

3.1 Differential Privacy

Central and local models. In the central model, a trusted curator collects raw user data and releases randomized outputs satisfying differential privacy (DP). In contrast, the local model eliminates the trusted curator by requiring each user to apply a local randomizer before sending data to the server, resulting in local differential privacy (LDP) guarantees.

Definition 1 (Differential Privacy [13,14])**.** A randomized mechanism $\mathcal{M}: \mathcal{X}^n \to \mathcal{Y}$ satisfies (ε,δ)-DP if, for any neighboring datasets $X, X' \in \mathcal{X}^n$ (differing in exactly one user's record) and any measurable set $Y \subseteq \mathcal{Y}$,

$$\Pr[\mathcal{M}(X) \in Y] \leq e^{\varepsilon} \Pr[\mathcal{M}(X') \in Y] + \delta.$$

We review two important theorems. The post-processing theorem states that an adversary cannot increase the privacy loss by analyzing the mechanism output without the help of additional knowledge of the private dataset.

Theorem 1. (Post-processing Theorem [14]) Let $\mathcal{M}: \mathcal{X}^n \to \mathcal{Y}$ be a randomized mechanism satisfying (ε,δ)-DP and $f: \mathcal{Y} \to \mathcal{Z}$ be any arbitrary randomized mapping. Then $f \circ M: \mathcal{X}^n \to \mathcal{Z}$ satisfies (ε,δ)-DP.

The composition theorem allows the computation of cumulative privacy loss for multiple queries on the raw dataset.

Theorem 2. (Composition Theorem [14]) Let $\mathcal{M}_1: \mathcal{X}^n \to \mathcal{Y}_1$ satisfies (ε_1,δ_1)- differential privacy and $\mathcal{M}_2: \mathcal{X}^n \to \mathcal{Y}_2$ satisfies (ε_2,δ_2)-DP. For any $X \in \mathcal{X}^n$, the mechanism $\mathcal{M} = (\mathcal{M}_1(X), \mathcal{M}_2(X))$ satisfies $(\varepsilon_1+\varepsilon_2, \delta_1+\delta_2)$-DP.

Shuffle model. The shuffle model of differential privacy lies between the local and the central models. A shuffle protocol consists of three components $\mathcal{P} = (\mathcal{R}, \mathcal{S}, \mathcal{A})$. Each user $i \in [n]$ applies a weak local randomizer $\mathcal{R}$ to their input $x_i \in \mathcal{X}$ and sends the resulting one or more messages to a shuffler $\mathcal{S}$. The shuffler applies a uniformly random permutation to all submitted messages and forwards only the resulting *multiset* to an analyzer $\mathcal{A}$ (implemented by the server), thereby removing the linkage between users and individual messages.

Definition 2 (Shuffle-Model DP [10])**.** A protocol $\mathcal{P} = (\mathcal{R}, \mathcal{S}, \mathcal{A})$ satisfies (ε,δ)-DP in the shuffle model if the mechanism

$$(x_1, \ldots, x_n) \mapsto \mathcal{S}\big(\mathcal{R}(x_1), \ldots, \mathcal{R}(x_n)\big)$$

is (ε,δ)-DP as a function from $\mathcal{X}^n$ to shuffled multisets.

To capture our protocols, we extend this model to allow each user to receive a data-independent auxiliary input prior to the local randomizer.

DEFINITION 3 (Shuffle Model with Data-Independent Auxiliary Input; this work)**.** The shuffle model with data-independent auxiliary input is parameterized by an auxiliary-input space $\mathcal{T}$ and a joint distribution $\mathcal{D}_{\text{aux}}$ over $\mathcal{T}^n$ that is independent of the input dataset. A protocol $\mathcal{P} = (\mathcal{R}, \mathcal{S}, \mathcal{A})$ satisfies (ε, δ)-DP in this model if the mechanism

$$(x_1, \ldots, x_n) \mapsto \mathcal{S}\big(\mathcal{R}(x_1, \theta_1), \ldots, \mathcal{R}(x_n, \theta_n)\big),$$

where $(\theta_1, \ldots, \theta_n) \sim \mathcal{D}_{\text{aux}}$, is (ε, δ)-DP as a function from $\mathcal{X}^n$ to shuffled multisets.

Under both the standard model (Def. 2) and our extended model (Def. 3), the shuffler $\mathcal{S}$ performs *no computation beyond permutation*, since it does not drop, insert, modify, or tag reported messages, and it reveals nothing besides the permuted multiset. We treat $\mathcal{S}$ as an ideal shuffler and assume it does not collude with $\mathcal{A}$. In this work, we study shuffle-model protocols for frequency estimation over a domain $\mathcal{X}$ with n users holding inputs $x_i \in \mathcal{X}$, where each user may send a single message or multiple messages under the protocol.

Our protocols realize $\mathcal{D}_{\text{aux}}$ through a data-independent preprocessing step that requires only the shuffler's permutation capability, used to deliver auxiliary inputs from the analyzer to the users. In the binary protocol, $\theta_i = b_i \in \{0,1\}$ is the mode flag and $\mathcal{D}_{\text{aux}}$ is a uniform permutation of $n/2$ copies of 0 and $n/2$ copies of 1; in the histogram protocol, $\theta_i = (j_i, b_i) \in [d] \times \{0,1\}$ specifies the assigned noise bin and the mode flag.

3.2 Poisoning Robustness

Poisoning attacks. We consider a poisoning adversary that corrupts m of n users. Corrupted users may arbitrarily deviate from the prescribed local randomizer. In particular, they may inject *crafted in-domain messages* that are syntactically valid (i.e., lie in the same message domain as honest reports) and therefore cannot be filtered by domain membership once anonymized by shuffling. Accordingly, poisoning influence is governed by how many in-domain messages each corrupted user can contribute per round, and how much each such message perturbs the target statistic once aggregated.

Per-user message-limit premise. All poisoning-influence bounds in this paper are stated under the standard *per-user message limit* premise: in each collection round, the analyzer accepts at most $k+1$ messages from any user (one report plus up to k noise messages), including corrupted users. This premise is essential for bounded-influence robustness in the standard shuffle model. Without an enforceable bound on accepted messages, an adversary could cause arbitrarily many in-domain messages to be accepted, making the poisoning influence unbounded by repetition. We discuss a simple enforcement mechanism as a deployment option in App. C.

Poisoning influence. We quantify poisoning attacks by the worst-case impact on the analyzer's output, measured as the maximum shift in the expected estimate (in ℓ_1 distance).

Let f denote the target statistic. In the binary case or for a single histogram bin, $f \in \mathbb{R}$ is a scalar frequency; for histogram estimation, $f \in \mathbb{R}^d$ is the full frequency vector. Let $\tilde{f}$ be the estimate produced when all users are honest, and let $\tilde{f}^{\mathsf{Adv}}$ be the resulting estimate under attack. We define the *poisoning influence* as

$$\text{Infl}(m) \;:=\; \sup_{\mathsf{Adv}} \big\| \mathbb{E}\big[\tilde{f}\big] - \mathbb{E}\big[\tilde{f}^{\mathsf{Adv}}\big] \big\|_1 ,$$

where the supremum ranges over all poisoning strategies corrupting m users subject to the per-user message limit $k+1$.

Illustrative example. To illustrate how poisoning arises in the shuffle model, consider binary mean estimation via randomized response. Each honest user holds $x \in \{0,1\}$ and reports y such that for some $p \in (1/2, 1)$, $\Pr[y = x] = p$ and $\Pr[y = 1-x] = 1-p$. Given n reports, the standard unbiased estimator is

$$\tilde{f} \;=\; \frac{1}{2p-1}\left(\frac{1}{n}\sum_{i=1}^{n} y_i - (1-p)\right).$$

For any honest user, $\mathbb{E}[y_i] = (2p-1)x_i + (1-p)$, so

$$\mathbb{E}[\tilde{f}] = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

If each corrupted user $i \in \{1, \ldots, m\}$ injects *one* in-domain bit $c_i \in \{0,1\}$ (e.g., always 1), then the poisoned estimator is

$$\tilde{f}^{\mathsf{Adv}} \;=\; \frac{1}{2p-1}\left(\frac{1}{n}\left(\sum_{i=1}^{m} c_i + \sum_{i=m+1}^{n} y_i\right) - (1-p)\right),$$

and the expectation is

$$\mathbb{E}[\tilde{f}^{\mathsf{Adv}}] \;=\; \frac{1}{n}\sum_{i=m+1}^{n} x_i \;+\; \frac{1}{2p-1}\cdot\frac{1}{n}\sum_{i=1}^{m}(c_i - (1-p)).$$

Therefore, the expected estimate is shifted by

$$\text{Infl}(m) \;=\; \frac{m}{n}\cdot\frac{p}{2p-1},$$

demonstrating that even a small number of corrupted users can exert non-negligible influence.

If a corrupted user could inject t accepted in-domain bits, the expected shift would scale linearly with t, underscoring the necessity of an enforceable per-user message limit.

Robustness. Under the standard per-user message limit, poisoning robustness is governed by how much a corrupted user can bias the estimator within this constrained action budget. Concretely, two protocol-induced factors determine $\text{Infl}(m)$: (i) the estimator's per-message effect on the target statistic once messages are aggregated, and (ii) the corrupted user's feasible in-domain message multiset. Intuitively, allowing

richer multi-message reporting can improve accuracy by enabling more expressive noise synthesis, but it may also enlarge the feasible multiset space and thus increase the adversary's degrees of freedom to introduce bias. Our protocol is designed to reconcile these goals, supporting accurate multi-message noise generation while restricting each user's multiset to a structured family, yielding explicit and tunable upper bounds on $\mathrm{Infl}(m)$ under the enforced message-limit premise.

4 Binary Frequency Estimation

We begin with binary frequency estimation (equivalently, binary summation). Each user $i \in [n]$ holds a bit $x_i \in \{0,1\}$, and the analyzer aims to estimate the fraction of ones

$$f = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

For convenience, we assume n is even; this assumption can be removed by rounding and affects only constants.

Design goal. We seek a shuffle-model protocol that simultaneously achieves: (i) (ε,δ)-DP, (ii) low estimation error, and (iii) an explicit bound on worst-case estimator bias (*poisoning influence*) under *poisoning attacks*, where corrupted users may deviate from the local randomizer but are restricted to sending only protocol-admissible in-domain messages (in particular, respecting the per-user message cap).

4.1 Protocol overview

Our binary protocol has the form $\mathcal{P}^B = (\mathcal{R}^B, \mathcal{S}, \mathcal{A}^B)$. The central idea is to realize a low-variance *symmetric binomial-sum noise* distribution while explicitly restricting each user's feasible anonymous message multiset.

Preprocessing: shuffled mode assignment. We introduce a *data-independent* preprocessing step that assigns each user a *mode flag* $b_i \in \{0,1\}$, hidden from the analyzer. Concretely, the analyzer submits to the shuffler a multiset containing $n/2$ copies of 0 and $n/2$ copies of 1. The shuffler uniformly permutes these mode flags, and each user i receives one permuted mode flag b_i as auxiliary input for the local randomizer; the analyzer never observes the per-user assignment. Intuitively, this commits each user to one of two complementary noise-sampling probabilities (p vs. $1-p$) prior to reporting.

Local randomizer. Given input x_i and the assigned mode flag b_i, user i samples one noise bit $\eta_i \leftarrow \mathrm{Ber}(p_{b_i})$, where

$$p_{b_i} = \begin{cases} p, & b_i = 0, \\ 1-p, & b_i = 1, \end{cases} \qquad \text{with } p \le \tfrac{1}{2}.$$

User i then sends $x_i + \eta_i$ copies of the in-domain message 1 to the shuffler. Thus each user sends either 0, 1, or 2 copies of message 1, and in particular at most 2 messages.

Analyzer. Let $|\mathbf{Y}|$ denote the total number of 1-messages received after shuffling. The analyzer outputs

$$\tilde{f} = \frac{|\mathbf{Y}|}{n} - \frac{1}{2}.$$

This is a *constant-shift* estimator.

Noise intuition. By construction, $n/2$ users receive $b_i = 0$ and $n/2$ receive $b_i = 1$. Hence the aggregate noise satisfies

$$\sum_{i=1}^{n} \eta_i \sim \mathrm{Bin}\left(\tfrac{n}{2}, p\right) + \mathrm{Bin}\left(\tfrac{n}{2}, 1-p\right), \quad \mathbb{E}\left[\sum_{i=1}^{n} \eta_i\right] = \tfrac{n}{2}.$$

Thus, $\tilde{f}$ is unbiased for f, and the analyzer applies only a constant shift rather than a multiplicative debiasing factor that could amplify the effect of adversarial message injections.

Communication complexity. Each user sends $x_i + \eta_i$ copies of message 1 and thus at most 2 messages. Moreover, since marginally $\mathbb{E}[\eta_i] = 1/2$ for each user under the balanced mode assignment, the expected number of noise messages is $n/2$. For the all-ones input ($x_i = 1$ for all i), the expected total number of transmitted messages is $n + n/2 = 1.5n$, i.e., 1.5 messages per user on average.

4.2 Algorithms

The local randomizer and analyzer are summarized in Algs. 1 and 2. The parameter p will be instantiated as a function of (n,ε,δ) in the privacy analysis.

Algorithm 1 $\mathcal{R}^B_{b,p}$: randomizer for the binary protocol

Input: $x \in \{0,1\}$, mode flag $b \in \{0,1\}$
Output: multiset $\mathbf{y}$ consisting of copies of symbol 1
if $b = 0$ **then**
    $p_b \leftarrow p$
else
    $p_b \leftarrow 1-p$
end if
Sample $\eta \leftarrow \mathrm{Ber}(p_b)$
Append $x+\eta$ copies of message 1 to $\mathbf{y}$
Return $\mathbf{y}$

Algorithm 2 $\mathcal{A}^B$: analyzer for the binary protocol

Input: shuffled multiset $\mathbf{Y}$ consisting of copies of 1 ▷ all users' messages after shuffling
Output: estimate $\tilde{f} \in \mathbb{R}$
$\tilde{f} \leftarrow \frac{|\mathbf{Y}|}{n} - \frac{1}{2}$
Return $\tilde{f}$

4.3 Main guarantee

We now state the main theorem for the binary protocol.

THEOREM 3. For any $\varepsilon \in (0,1]$, $\delta \in (0,1)$, and $n > \frac{60}{\varepsilon^2}\log\frac{4}{\delta}$, there exists a choice of parameter $p \in (0,1/2]$ such that the protocol $\mathcal{P}^B = (\mathcal{R}^B_{b,p}, \mathcal{S}, \mathcal{A}^B)$ satisfies:

1. **Privacy.** $\mathcal{P}^B$ satisfies (ε,δ)-DP.
2. **Accuracy.** For any $\beta \in (0,1)$, with probability at least $1-\beta$ over the protocol randomness, the error is $\left|\tilde{f} - f\right| = O\left(\frac{1}{\varepsilon n}\sqrt{\log\frac{1}{\delta}\cdot\log\frac{1}{\beta}}\right)$.
3. **Robustness.** Under any poisoning attack corrupting m users, the expected estimate can change by at most $\frac{3m}{2n}$.
4. **Communication.** Each user sends at most 2 messages, and for the all-ones input the expected number of messages is 1.5 per user.

4.4 Privacy Analysis

We prove (ε,δ)-DP by a reduction to a one-dimensional additive-noise mechanism, followed by post-processing [14].

LEMMA 1 (Reduction to a one-bit additive-noise mechanism)**.** If $\mathcal{C}^B_{n,p}$ in Alg. 3 satisfies (ε,δ)-DP, then the shuffle protocol $\mathcal{P}^B = (\mathcal{R}^B_{b,p}, \mathcal{S}, \mathcal{A}^B)$ satisfies (ε,δ)-DP.

Algorithm 3 $\mathcal{C}^B_{n,p}$ (core additive-noise counting mechanism)

Input: $x \in \{0,1\}$
Output: $y \in \mathbb{N}$
$z \leftarrow 0$
for all $i \in [n/2]$ **do**
  Sample $z_1 \sim \mathrm{Ber}(p)$
  Sample $z_2 \sim \mathrm{Ber}(1-p)$
  $z \leftarrow z + z_1 + z_2$
end for
$y \leftarrow x + z$
Return y

Proof. The output is $\mathcal{M}(\mathbf{x}) = \mathcal{S}\big(\mathcal{R}^B_{b_1,p}(x_1), \ldots, \mathcal{R}^B_{b_n,p}(x_n)\big)$, where the mode-flag assignment $(b_1,\ldots,b_n)$, a uniform permutation of $n/2$ copies of 0 and $n/2$ copies of 1 sampled independently of the user data $\mathbf{x}$, serves as the auxiliary input of Def. 3. Under the non-collusion assumption, this assignment is not part of this output, so it suffices to characterize the distribution of the shuffled multiset. Fix any such assignment; given $\mathbf{x} = (x_1,\ldots,x_n) \in \{0,1\}^n$, user i samples $\eta_i \sim \mathrm{Ber}(p_{b_i})$ independently and outputs a multiset of $x_i + \eta_i$ copies of the same symbol 1.

After shuffling, the analyzer observes only the multiset of all messages. Since all shuffled messages are identical, this multiset is fully determined by the message count $|\mathbf{Y}|$,

$$T(\mathbf{x}) := |\mathbf{Y}| = \sum_{i=1}^{n}(x_i + \eta_i) = \left(\sum_{i=1}^{n} x_i\right) + Z,$$

where $Z := \sum_{i=1}^{n}\eta_i$ depends only on protocol randomness and is independent of $\mathbf{x}$. Any permutation applied by the shuffler is thus immaterial.

To verify DP, it suffices to prove that $T(\cdot)$ is (ε,δ)-DP, because the full protocol output (including the shuffled multiset and the analyzer output) is a deterministic post-processing of T. Indeed, $\mathcal{A}^B$ depends only on $|\mathbf{Y}| = T(\mathbf{x})$.

Now fix any neighboring datasets $\mathbf{x}, \mathbf{x}'$ that differ in exactly one record, say user j. Let $c := \sum_{i\neq j} x_i = \sum_{i\neq j} x'_i$, which is a constant for this neighboring pair. Then

$$T(\mathbf{x}) = c + (x_j + Z), \qquad T(\mathbf{x}') = c + (x'_j + Z).$$

Since adding a constant c is post-processing, it suffices to show that the one-dimensional mechanism $x \mapsto x + Z$ is (ε,δ)-DP for $x \in \{0,1\}$.

Finally, by grouping the n Bernoulli trials into $n/2$ pairs with parameters p and $1-p$, we can write $Z = \sum_{i=1}^{n/2} z_i$ where $z_i \sim \mathrm{Ber}(p) + \mathrm{Ber}(1-p)$ are i.i.d. Thus the mechanism $x \mapsto x + Z$ is exactly $\mathcal{C}^B_{n,p}$ up to post-processing. Therefore, if $\mathcal{C}^B_{n,p}$ is (ε,δ)-DP, then so is $T(\cdot)$, and hence $\mathcal{P}^B$ is (ε,δ)-DP by the post-processing theorem (Theorem 1). □

By Lemma 1, it remains to prove that $\mathcal{C}^B_{n,p}$ satisfies (ε,δ)-DP for a suitable choice of p.

THEOREM 4 (Privacy of the Symmetric Binomial-Sum Mechanism)**.** For $\varepsilon > 0$, $\delta \in (0,1)$, and $n \geq \frac{2L(3+2a)}{3a^2}$, where $a = \tanh\frac{\varepsilon}{2}$ and $L = \log\frac{4}{\delta}$, there exists a choice of parameter $p \in (0,1/2]$ such that $\mathcal{C}^B_{n,p}$ (Alg. 3) satisfies (ε,δ)-DP.

Proof. Recall that $\mathcal{C}^B_{n,p}(x) = x + Z$, where $Z = \sum_{i=1}^{n/2} z_i$, $z_i \sim \mathrm{Ber}(p) + \mathrm{Ber}(1-p)$, independently. Since the input domain is $\{0,1\}$, it suffices to verify DP for the neighboring inputs $x = 0$ and $x = 1$, i.e., for all measurable $Y \subseteq \mathbb{N}$,

$$\begin{aligned}\Pr[\,\mathcal{C}^B_{n,p}(1) \in Y\,] &\leq e^{\varepsilon}\Pr[\,\mathcal{C}^B_{n,p}(0) \in Y\,] + \delta,\\ \Pr[\,\mathcal{C}^B_{n,p}(0) \in Y\,] &\leq e^{\varepsilon}\Pr[\,\mathcal{C}^B_{n,p}(1) \in Y\,] + \delta. \qquad (1)\end{aligned}$$

Step 1: Tail bound. The noise distribution is symmetric with mean $\mathbb{E}[Z] = n/2$. Let $\overline{Z} := Z - n/2$ and $\bar{z}_i := z_i - 1$. Then $\overline{Z} = \sum_{i=1}^{n/2}\bar{z}_i$, where each $\bar{z}_i$ is zero-mean, bounded in $[-1,1]$, and has distribution

$$\Pr[\bar{z} = y] = \begin{cases} p_0 & y = -1,\\ 1-2p_0 & y = 0,\\ p_0 & y = 1,\end{cases} \qquad \text{with } p_0 = p(1-p).$$

Moreover, $\mathbb{E}[\bar{z}^2] = 2p_0$, hence $\sum_{i=1}^{n/2}\mathbb{E}[\bar{z}_i^2] = (n/2)\cdot 2p_0 = np_0$.

Fact 1 (Bernstein inequality)**.** Let $x_1,\ldots,x_m$ be independent zero-mean random variables with $|x_i| \leq M$ for all i. Then for all $t > 0$,

$$\Pr\left[\sum_{i=1}^{m} x_i \geq t\right] \leq \exp\left(-\frac{t^2/2}{\sum_{i=1}^{m}\mathbb{E}[x_i^2] + Mt/3}\right).$$

Applying Fact 1 to $\overline{Z}$ with $m = n/2$, $M = 1$, and $\sum \mathbb{E}[\overline{z}_i^2] = np_0$, we obtain

$$\Pr[\overline{Z} \geq t] \leq \exp\left(-\frac{t^2/2}{np_0 + t/3}\right).$$

By symmetry, this yields

$$\Pr[|Z - n/2| \geq t] \leq 2\exp\left(-\frac{t^2/2}{np(1-p) + t/3}\right). \quad (2)$$

For a target tail limit $\delta_t \in (0,1)$, define

$$t = \left\lceil \frac{1}{3}\log\frac{2}{\delta_t} + \sqrt{\left(\frac{1}{3}\log\frac{2}{\delta_t}\right)^2 + 2np(1-p)\log\frac{2}{\delta_t}} \right\rceil. \quad (3)$$

Then Eq. (2) implies

$$\Pr[|Z - n/2| > t] \leq \delta_t. \quad (4)$$

Step 2: Likelihood-ratio bound on the central region. Define the "good" (central) set

$$G := \{k \in \mathbb{N} : |k - n/2| \leq t\}.$$

For $k \in \mathbb{N}$, the DP inequality for the singleton event $\{k\}$ reads

$$\Pr[Z = k-1] \leq e^{\varepsilon}\Pr[Z = k] + \delta(\{k\}),$$

since $\mathcal{C}^B_{n,p}(1) = 1 + Z$ and $\mathcal{C}^B_{n,p}(0) = Z$. Thus, a sufficient condition for Eq. (1) is: (i) the tail probability outside G is at most δ_t, and (ii) for all k with $k, k-1 \in G$,

$$\frac{\Pr[Z = k]}{\Pr[Z = k-1]} \leq e^{\varepsilon_t} \quad \text{for some } \varepsilon_t \leq \varepsilon. \quad (5)$$

Since Z is symmetric about $n/2$, i.e., $\Pr[Z = k] = \Pr[Z = n-k]$, the involution $k \mapsto n+1-k$ maps the adjacent-pair range $G^* := \{k : k, k-1 \in G\}$ to itself and turns each ratio $\Pr[Z = k]/\Pr[Z = k-1]$ into its reciprocal. Hence the ratio bound on G^* controls both directions of the singleton DP inequalities; the two directions are assembled in Step 3.

To establish Eq. (5), we use the following monotonicity lemma (proved in the full version of this paper).

Lemma 2 (Monotone likelihood ratio of the symmetric Bernoulli sum)**.** Let $p \in (0, 1/2]$, and for Bernoulli distributions

$$\xi \sim \begin{bmatrix} 0 & 1 \\ p & 1-p \end{bmatrix}, \qquad \eta \sim \begin{bmatrix} 0 & 1 \\ 1-p & p \end{bmatrix},$$

let $X = \sum_{i=1}^N \xi_i + \sum_{j=1}^N \eta_j$, where all summands are independent. Then the ratio $\Pr[X = m+1]/\Pr[X = m]$ is non-increasing in m over all m with $\Pr[X = m] > 0$ and $\Pr[X = m+1] > 0$.

In our setting, Z is a sum of n Bernoulli trials: $n/2$ with parameter p and $n/2$ with parameter $1-p$. Pairing them yields i.i.d. summands $z_i = \mathrm{Ber}(p) + \mathrm{Ber}(1-p)$; equivalently, Z has the same distribution as the variable X in Lemma 2 with $N = n/2$. Hence, $\Pr[Z = m+1]/\Pr[Z = m]$ is non-increasing in m over its support.

Therefore, for every k with $k, k-1 \in G$ and $k \leq n/2$,

$$\frac{\Pr[Z = k]}{\Pr[Z = k-1]} \leq \frac{\Pr[Z = n/2 - t + 1]}{\Pr[Z = n/2 - t]},$$

and by symmetry the worst-case in-G ratio is attained at $k = n/2 - t + 1$. Define

$$\varepsilon_t := \log\left(\frac{\Pr[Z = n/2 - t + 1]}{\Pr[Z = n/2 - t]}\right). \quad (6)$$

Then Eq. (5) holds with this ε_t.

Step 3: The DP inequality. Let $Y \subseteq \mathbb{N}$ be arbitrary. On the central region, the ratio bound Eq. (5) covers every k with $k, k-1 \in G$, that is, $k \in [n/2 - t + 1, n/2 + t]$. For the shifted count $1 + Z$, the only uncovered point of $G + 1 = [n/2 - t + 1, n/2 + t + 1]$ is the right boundary $k = n/2 + t + 1$. Its mass $\Pr[1 + Z = n/2 + t + 1] = \Pr[Z = n/2 + t]$ is at most $\delta_t/2$ by the symmetric tail Eq. (2) (using $\Pr[Z \geq n/2 + t] = \frac{1}{2}\Pr[|Z - n/2| \geq t]$). Together with the tail bound Eq. (4), this gives

$$\begin{aligned}\Pr[1 + Z \in Y] &\leq e^{\varepsilon_t}\Pr[Z \in Y] + \Pr[Z \notin G] + \Pr[Z = n/2 + t] \\ &\leq e^{\varepsilon_t}\Pr[Z \in Y] + \tfrac{3}{2}\delta_t.\end{aligned}$$

The reverse direction $\Pr[Z \in Y] \leq e^{\varepsilon_t}\Pr[1 + Z \in Y] + \frac{3}{2}\delta_t$ follows by applying the above to $Y'' := n + 1 - Y$ and using $\Pr[Z = k] = \Pr[Z = n - k]$, which swaps $\Pr[Z \in Y]$ with $\Pr[1 + Z \in Y]$. Thus, $\mathcal{C}^B_{n,p}$ satisfies $(\varepsilon_t, \frac{3}{2}\delta_t)$-DP. In particular, if we choose parameters so that $\varepsilon_t \leq \varepsilon$ and $\frac{3}{2}\delta_t \leq \delta$, then $\mathcal{C}^B_{n,p}$ is (ε, δ)-DP.

Step 4: Existence of a feasible choice of p and the condition on n. Eq. (3) and Eq. (6) characterize $(\varepsilon_t, \delta_t)$ as functions of (n, p). Intuitively, increasing the noise dispersion (achieved by taking p closer to $1/2$) decreases the worst-case adjacent ratio and hence decreases ε_t. In particular, taking $p = 1/2$ gives $Z \sim \mathrm{Bin}(n, 1/2)$, and then

$$\varepsilon_t = \log\left(\frac{n/2 + t}{n/2 - t + 1}\right).$$

Choosing $\delta_t = \delta/2$ (so that $\frac{3}{2}\delta_t = \frac{3}{4}\delta \leq \delta$ per Step 3, aligning $L = \log\frac{2}{\delta_t} = \log\frac{4}{\delta}$ with the constant used in Theorems 3 and 5) and t as in Eq. (3) (with $p = 1/2$), it suffices to ensure

$$\frac{n/2 + t}{n/2 - t + 1} \leq e^{\varepsilon}.$$

Ignoring the ceiling and the $+1$ constant, consider the relaxed condition

$$\frac{n/2 + t}{n/2 - t} \leq e^{\varepsilon}.$$

This inequality is equivalent to

$$t \le \frac{n}{2}\tanh\frac{\varepsilon}{2}. \tag{7}$$

Substituting the expression for t from Eq. (3) (with $p = 1/2$) into Eq. (7), we require

$$\frac{1}{3}\log\frac{4}{\delta} + \sqrt{\left(\frac{1}{3}\log\frac{4}{\delta}\right)^2 + \frac{n}{2}\log\frac{4}{\delta}} \le \frac{n}{2}\tanh\frac{\varepsilon}{2}. \tag{8}$$

Let $a = \tanh(\varepsilon/2)$ and $L = \log\frac{4}{\delta}$. Isolating the square root in Eq. (8), squaring (valid since $na/2 - L/3 \ge 0$ for $n \ge 2L/(3a)$, implied by the bound below), and rearranging yields the sufficient condition

$$n \ge \frac{2L(3+2a)}{3a^2}. \tag{9}$$

This is exactly the lower bound stated in the theorem. Therefore, when n satisfies this condition, choosing $p = 1/2$ ensures $\varepsilon_t \le \varepsilon$ and $\delta_t \le \delta$, and the mechanism satisfies (ε,δ)-DP. For larger n, one can choose $p < 1/2$ appropriately, so a suitable $p \in (0, 1/2]$ always exists when n meets the stated bound. As $a \to 0$ (small ε), the bound asymptotes to $n \ge \frac{2L}{a^2} \approx \frac{8}{\varepsilon^2}\log\frac{4}{\delta}$. □

4.5 Accuracy Analysis

This section analyzes the accuracy of our binary protocol.

Claim 1. Fix any $p \in (0, 1/2]$ and $n \in \mathbb{N}$. For any raw data $x_1, \dots, x_n \in \{0,1\}$, the error variance of protocol $\mathcal{P}^B = (\mathcal{R}^B_{b,p}, \mathcal{S}, \mathcal{A}^B)$ is $p(1-p)/n$.

Proof. Each user contributes an independent Bernoulli noise with variance $p(1-p)$. Since the analyzer outputs the average over n users, the variance of the aggregated estimate is $1/n^2 \cdot n \cdot p(1-p) = p(1-p)/n$. □

By standard Bernstein-type concentration bounds (Fact 1 and Eq. (3)), for any $\beta \in (0,1)$, with probability at least $1-\beta$, the estimation error of $\mathcal{P}^B$ satisfies

$$\left|\tilde{f} - f\right| \le \frac{1}{3n}\log\frac{2}{\beta} + \frac{1}{n}\sqrt{\left(\frac{1}{3}\log\frac{2}{\beta}\right)^2 + 2np(1-p)\log\frac{2}{\beta}}. \tag{10}$$

The privacy of $\mathcal{P}^B$ is established by Theorem 4 above via direct Bernstein analysis on $\mathcal{C}^B_{n,p}$, giving the tightest closed-form sample size $n = O(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$ in the small-ε regime. It does not, however, pin down an explicit value of the noise parameter p as a function of (ε,δ,n), which the high-probability error bound above requires.

To obtain such an explicit p, we introduce an auxiliary mechanism $\mathcal{C}^B_{n,b,p}$ (Alg. 4) that exposes additional structure, with its output (y_1, y_2) reporting the per-group 1-counts of mode-0 and mode-1 users separately, rather than only their sum. Since the message count of $\mathcal{P}^B$ equals $y_1 + y_2$, the protocol $\mathcal{P}^B$ is a deterministic post-processing of this auxiliary mechanism. By the post-processing inequality, any (ε,δ)-DP guarantee therefore transfers to $\mathcal{P}^B$ (Claim 2); this transfer is a self-contained but looser alternative to Theorem 4. Since post-processing cannot improve accuracy, the accuracy of $\mathcal{P}^B$ is no worse than that of the auxiliary mechanism. Theorem 5 below provides this p; substituting it into the high-probability bound above then yields explicit accuracy guarantees for $\mathcal{P}^B$ in terms of (ε,δ).

Algorithm 4 $\mathcal{C}^B_{n,b,p}$

1: **Input:** $x_1, \dots, x_n \in \{0,1\}$, mode flags $(b_1, \dots, b_n)$
2: **Output:** $y_1, y_2 \in \mathbb{N}$
3: $y_1 = \sum_{i=1}^n \mathbb{I}(b_i)x_i + Bin(\frac{n}{2}, p)$
4: $y_2 = \sum_{i=1}^n \mathbb{I}(1-b_i)x_i + Bin(\frac{n}{2}, 1-p)$
5: **Return:** (y_1, y_2)

Lemma 3 (Likelihood-Ratio Inequalities)**.** Fix $\varepsilon \in (0,1]$, $p \in (0, 2/5]$, and set $\alpha = \varepsilon/2$. For any real $\mu \ge \frac{12\log 4}{\varepsilon^2}$ (instantiated as the noise-group mean $\mu = np/2$ in Theorem 5), define

$$F_+ = \frac{1}{2}\left[\frac{(1-p)(1+\alpha+1/\mu)}{1-p(1+\alpha)} + \frac{1-p(1-\alpha)+p/\mu}{(1-p)(1-\alpha)}\right],$$
$$F_- = \frac{1}{2}\left[\frac{(1-p)(1-\alpha)}{1-p(1-\alpha)+p/\mu} + \frac{1-p(1+\alpha)}{(1-p)(1+\alpha+1/\mu)}\right].$$

Then $F_+ \le e^{\varepsilon}$ and $F_- \ge e^{-\varepsilon}$. The proof is given in the full version of this paper.

Theorem 5. For any $\varepsilon,\delta \in (0,1]$ and $n \ge \frac{60}{\varepsilon^2}\log\frac{4}{\delta}$, $\mathcal{C}^B_{n,b,p}$ (Alg. 4) with $p = \frac{24}{\varepsilon^2 n}\log\frac{4}{\delta}$ satisfies (ε,δ)-DP.

Proof. Let $n_0 := n/2$, $L := \log\frac{4}{\delta}$, $\alpha := \varepsilon/2$, and $\mu := n_0 p = \frac{12L}{\varepsilon^2}$. Since $L \ge \log 4$ for $\delta \le 1$, the given conditions imply $\mu \ge \frac{12\log 4}{\varepsilon^2}$; since $n \ge \frac{60L}{\varepsilon^2}$, they imply $p \le 2/5$. The hypotheses of Lemma 3 are therefore met.

Likelihood-ratio formula. We compute the per-output likelihood ratio $R(y_1,y_2)$ explicitly. Let $\xi \sim \mathrm{Bin}(n_0, p)$ and $\eta \sim \mathrm{Bin}(n_0, 1-p)$ be independent, and set $\bar{\eta} := n_0 - \eta \sim \mathrm{Bin}(n_0, p)$. For any $(y_1,y_2) \in \mathbb{N}^2$, the two input cases decompose as

$$\Pr[\mathcal{C}^B_{n,b,p}(0) = (y_1,y_2)] = \Pr[\xi = y_1]\Pr[\eta = y_2],$$
$$\Pr[\mathcal{C}^B_{n,b,p}(1) = (y_1,y_2)] = \frac{1}{2}\big(\Pr[\xi = y_1 - 1]\Pr[\eta = y_2] + \Pr[\xi = y_1]\Pr[\eta = y_2 - 1]\big),$$

where the input-1 case splits into two equiprobable sub-cases according to which group receives the extra count. Applying the binomial ratio identity $\frac{\Pr[\xi=y_1-1]}{\Pr[\xi=y_1]} = \frac{1-p}{p}\cdot\frac{y_1}{n_0-y_1+1}$ and the analogous identity for η, we obtain

$$R(y_1,y_2) := \frac{\Pr[\mathcal{C}^B_{n,b,p}(1) = (y_1,y_2)]}{\Pr[\mathcal{C}^B_{n,b,p}(0) = (y_1,y_2)]} \tag{11}$$

$$= \frac{1}{2}\left[\frac{1-p}{p}\cdot\frac{y_1}{n_0-y_1+1}+\frac{p}{1-p}\cdot\frac{y_2}{n_0-y_2+1}\right].$$

Good events. We split the privacy analysis into a typical (high-probability) region, where the binomial counts y_1 and n_0-y_2 stay near their common mean μ, and the complementary bad event, whose probability we bound by δ via Chernoff. Define the typical regions

$$G_+ := \{y_1 \le \mu(1+\alpha)+1,\ n_0-y_2+1 \ge \mu(1-\alpha)\},$$
$$G_- := \{y_1 \ge \mu(1-\alpha),\ n_0-y_2 \le \mu(1+\alpha)\}.$$

G_+ controls the upper tails relevant for the forward direction, where the $+1$ buffer in y_1 absorbs the input-1 shift; G_- controls the symmetric opposite tails for the reverse direction.

Forward direction. We bound the LR on G_+ by bounding each of the two summands in Eq. (11). The constraints $y_1 \le \mu(1+\alpha)+1$ and $n_0-y_1+1 \ge n_0-\mu(1+\alpha)$ from G_+ give

$$\frac{1-p}{p}\cdot\frac{y_1}{n_0-y_1+1} \le \frac{(1-p)(\mu(1+\alpha)+1)}{p\,(n_0-\mu(1+\alpha))} = \frac{(1-p)(1+\alpha+1/\mu)}{1-p(1+\alpha)};$$

analogously, the constraints $y_2 \le n_0-\mu(1-\alpha)+1$ and $n_0-y_2+1 \ge \mu(1-\alpha)$ from G_+ give

$$\frac{p}{1-p}\cdot\frac{y_2}{n_0-y_2+1} \le \frac{1-p(1-\alpha)+p/\mu}{(1-p)(1-\alpha)}.$$

Substituting both bounds into Eq. (11) yields $R(y_1,y_2) \le F_+(\varepsilon,p,\mu) \le e^{\varepsilon}$ on G_+, where the final inequality is Lemma 3.

It remains to bound the probability of the complementary bad event G_+^c under input 1. Multiplicative Chernoff applied to the two violation events gives

$$\Pr[y_1 > \mu(1+\alpha)+1] \le \Pr[\xi > \mu(1+\alpha)] \le \exp(-\alpha^2\mu/3) = e^{-L} = \delta/4,$$
$$\Pr[n_0-y_2+1 < \mu(1-\alpha)] \le \Pr[\bar{\eta} < \mu(1-\alpha)] \le \exp(-\alpha^2\mu/2) = e^{-3L/2} \le \delta/4.$$

A union bound therefore yields $\Pr[G_+^c \mid P_1] \le \delta/2 < \delta$, and combining the LR bound on G_+ with this tail bound on G_+^c via the standard good-event/bad-event argument gives the forward DP inequality

$$\Pr[\mathcal{C}^B_{n,b,p}(1) \in Y] \le e^{\varepsilon}\Pr[\mathcal{C}^B_{n,b,p}(0) \in Y] + \delta.$$

Reverse direction. A symmetric derivation on G_- that uses $y_1 \ge \mu(1-\alpha)$ and $n_0-y_2 \le \mu(1+\alpha)$ yields $R(y_1,y_2) \ge F_-(\varepsilon,p,\mu) \ge e^{-\varepsilon}$ by Lemma 3, and symmetric Chernoff bounds give $\Pr[G_-^c \mid P_0] \le \delta/2 < \delta$, so

$$\Pr[\mathcal{C}^B_{n,b,p}(0) \in Y] \le e^{\varepsilon}\Pr[\mathcal{C}^B_{n,b,p}(1) \in Y] + \delta.$$

The two directions together establish (ε,δ)-DP. □

Claim 2. If $\mathcal{C}^B_{n,b,p}$ (Alg. 4) satisfies (ε,δ)-DP, then $\mathcal{P}^B$ satisfies (ε,δ)-DP and is at least as accurate as $\mathcal{C}^B_{n,b,p}$.

Proof. The protocol $\mathcal{P}^B$ is obtained by post-processing the output of $\mathcal{C}^B_{n,b,p}$ by aggregating y_1+y_2. Post-processing preserves differential privacy and cannot increase the estimation error, which proves the claim. □

Substituting the explicit p from Theorem 5 into the high-probability bound above, via the accuracy transfer of Claim 2, yields the error bound for $\mathcal{P}^B$ stated in Theorem 6 below.

Theorem 6. Under the hypotheses of Theorem 5, with probability at least $1-\beta$, the estimation error of $\mathcal{P}^B$ satisfies

$$\left|\tilde{f}-\tfrac{1}{n}\textstyle\sum x_i\right| \le \frac{1}{3n}\log\frac{2}{\beta}+\frac{1}{n}\sqrt{\left(\frac{1}{3}\log\frac{2}{\beta}\right)^2+2np\log\frac{2}{\beta}}$$
$$\le \frac{1}{3n}\log\frac{2}{\beta}+\frac{1}{n}\sqrt{\left(\frac{1}{3}\log\frac{2}{\beta}\right)^2+\frac{48}{\varepsilon^2}\log\frac{4}{\delta}\log\frac{2}{\beta}}$$
$$= O\left(\frac{1}{\varepsilon n}\sqrt{\log\frac{1}{\delta}\cdot\log\frac{1}{\beta}}\right).$$

4.6 Robustness Analysis

We analyze robustness of the binary protocol against a poisoning adversary that corrupts m out of n users. We assume the standard *per-user message limit* holds even for corrupted users, so each user can submit at most 2 messages to the shuffler. (If an adversary could violate this limit, robustness would be unbounded without additional enforcement; see App. C.)

Estimator as a function of the total message count. Recall that every message is the same symbol 1, so the shuffled transcript is fully determined by its cardinality $M := |\mathbf{Y}|$. The estimator can be written as

$$\tilde{f} = \frac{|\mathbf{Y}|}{n}-\frac{1}{2} = \frac{M}{n}-\frac{1}{2}.$$

Therefore, for any attack strategy,

$$\left|\mathbb{E}[\tilde{f}]-\mathbb{E}[\tilde{f}^{\mathsf{Adv}}]\right| = \frac{1}{n}\left|\mathbb{E}[M]-\mathbb{E}[M^{\mathsf{Adv}}]\right|. \tag{12}$$

Decomposing honest and corrupted contributions. Let H be the set of honest users and C the set of corrupted users, with $|C|=m$ and $|H|=n-m$. Decompose the total message count in the honest execution as

$$M = \sum_{i\in H}(x_i+\eta_i) + \sum_{i\in C}(x_i+\eta_i),$$

where $\eta_i \in \{0,1\}$ is the Bernoulli noise bit (drawn according to the protocol). Under a poisoning attack, corrupted users may deviate from the local randomizer but must respect the per-user message limit. Let $t_i \in \{0,1,2\}$ denote the (possibly

adversarial) number of messages sent by corrupted user $i \in C$. Then the attacked execution has total count

$$M^{\mathsf{Adv}} \;=\; \sum_{i\in H}(x_i+\eta_i) \;+\; \sum_{i\in C} t_i.$$

Since the adversary does not affect honest users' local randomness, the honest contribution cancels in expectation:

$$\mathbb{E}[M]-\mathbb{E}[M^{\mathsf{Adv}}] \;=\; \mathbb{E}\left[\sum_{i\in C}(x_i+\eta_i)\right]-\mathbb{E}\left[\sum_{i\in C} t_i\right]. \tag{13}$$

Bounding poisoning influence. For each corrupted user $i \in C$, we have $t_i \in \{0,1,2\}$ deterministically, hence

$$0 \;\le\; \sum_{i\in C} t_i \;\le\; 2m \quad\Rightarrow\quad 0 \;\le\; \mathbb{E}\left[\sum_{i\in C} t_i\right] \;\le\; 2m. \tag{14}$$

In the honest execution, each corrupted user $i \in C$ would contribute $x_i+\eta_i$ messages, where η_i is a Bernoulli bit whose expectation equals $1/2$ (because the preprocessing assigns $b_i \in \{0,1\}$ with equal marginal probability and uses $p_{b_i} \in \{p, 1-p\}$, yielding $\mathbb{E}[\eta_i] = 1/2$ for every i). Therefore,

$$\mathbb{E}\left[\sum_{i\in C}(x_i+\eta_i)\right] \;=\; \sum_{i\in C} x_i \;+\; \frac{m}{2}. \tag{15}$$

Combining Eq. (13), Eq. (14), and Eq. (15),

$$\left|\mathbb{E}[M]-\mathbb{E}[M^{\mathsf{Adv}}]\right| = \left|\left(\sum_{i\in C} x_i+\frac{m}{2}\right)-\mathbb{E}\left[\sum_{i\in C} t_i\right]\right|.$$

Since $\sum_{i\in C} x_i \in [0,m]$ and $\mathbb{E}[\sum_{i\in C} t_i] \in [0,2m]$, we obtain

$$\left|\mathbb{E}[M]-\mathbb{E}[M^{\mathsf{Adv}}]\right| \le$$
$$\max\left\{\left(\sum_{i\in C} x_i+\frac{m}{2}\right)-0,\; 2m-\left(\sum_{i\in C} x_i+\frac{m}{2}\right)\right\} = \frac{3m}{2}.$$

Plugging into Eq. (12) yields

$$\left|\mathbb{E}[\tilde{f}]-\mathbb{E}[\tilde{f}^{\mathsf{Adv}}]\right| \le \frac{3m}{2n}. \tag{16}$$

Remark 1 (Flag-free split-trial alternative)**.** $\mathcal{P}^B$ relies on the shuffler delivering the data-independent flag $b_i \in \{0,1\}$ to user i. When this preprocessing step is unavailable, an alternative protocol that fits the standard shuffle model (no auxiliary input) is the *flag-free split-trial*: each user i locally samples two independent noise bits $\eta_i^{(0)} \sim \mathrm{Ber}(p)$ and $\eta_i^{(1)} \sim \mathrm{Ber}(1-p)$ and sends $x_i+\eta_i^{(0)}+\eta_i^{(1)}$ copies of the message 1. The aggregate noise $\mathrm{Bin}(n,p)+\mathrm{Bin}(n,1-p)$ preserves the symmetric binomial-sum structure of $\mathcal{P}^B$, so the likelihood-ratio template of Theorem 5 carries over at effective sample size $2n$. The per-user message cap rises from 2 to 3, the all-ones expected number of messages per user from 1.5 to 2, and the worst-case poisoning influence loosens from $3m/(2n)$ to $2m/n$. Therefore, $\mathcal{P}^B$ remains preferable for its tighter communication and influence bounds; the flag-free split-trial serves as a fallback.

5 Histogram Protocol

We now consider frequency estimation over a finite domain $[d]$. Each user $i \in [n]$ holds $x_i \in [d]$, and the analyzer estimates the histogram $f \in [0,1]^d$, where $f_j = \frac{1}{n}\sum_{i=1}^n \mathbf{1}[x_i = j]$. As in the binary case, our designs achieve low variance via a balanced mode assignment (half using p and half using $1-p$).

A naïve histogram extension. As a baseline, one can treat the histogram as d independent binary frequency estimation and run the binary protocol for each bin in parallel, where user i contributes to bin j as if holding the bit $\mathbf{1}[x_i = j]$. This inherits the per-bin privacy/accuracy/robustness guarantees of the binary protocol. The drawback is communication, as each user must locally generate noise for every bin, yielding $O(d)$ messages per user. We therefore seek protocols with communication independent of d (up to $\log d$ bits per message).

5.1 Low-Communication Histogram Protocol

We present a low-communication histogram protocol that reduces per-user communication from the naïve $O(d)$ to $O(k)$ messages, with k typically much smaller than d. The key idea is to use a data-independent preprocessing step to *bind each honest user* to a single noise bin, so that noise is injected in aggregate across users rather than per user across all bins.

Preprocessing (shuffled bin & mode assignment). The analyzer submits to the shuffler a multiset of pairs $(j,b) \in [d]\times\{0,1\}$ such that: (i) each bin j receives the same number of assignments, and (ii) within each bin, half the assignments have $b=0$ and half have $b=1$ (up to rounding). The shuffler uniformly permutes these pairs and delivers one permuted pair (j_i,b_i) to each user as auxiliary input; the analyzer never observes the per-user assignment.

Algorithm 5 $\mathcal{R}^{H1}_{b,j,p}$: randomizer for low-comm. histogram

Input: $x \in [d]$, assigned pair $(j,b) \in [d]\times\{0,1\}$
Output: multiset $\mathbf{y}$ of elements from $[d]$
Append x to $\mathbf{y}$ ▷ raw report
if $b = 0$ **then**
  $p_b \leftarrow p$
else
  $p_b \leftarrow 1-p$
end if
for $t \leftarrow 1$ **to** k **do**
  if $\mathrm{Ber}(p_b) = 1$ **then**
    Append j to $\mathbf{y}$ ▷ noise for assigned bin
  end if
end for
Return multiset $\mathbf{y}$ of elements from $[d]$

Algorithm 6 $\mathcal{A}^{H1}$: analyzer for low-comm. histogram

Input: shuffled multiset $\mathbf{Y}$ of elements in $[d]$ ▷ all users' messages after shuffling
Output: estimate $\tilde{f} \in \mathbb{R}^d$
Initialize counts $C_1, \ldots, C_d \leftarrow 0$
for all $y \in \mathbf{Y}$ **do**
    $C_y \leftarrow C_y + 1$
end for
for $j \leftarrow 1$ **to** d **do**
    $\tilde{f}_j \leftarrow \frac{C_j}{n} - \frac{k}{2d}$
end for
Return $\tilde{f}$

Local randomizer. (Alg. 5) User i sends one *raw* message equal to x_i. In addition, user i runs k independent Bernoulli trials with probability $p_{b_i} \in \{p, 1-p\}$ and, on success, sends a *noise* message equal to the assigned bin j_i. Thus, each user sends at most $k+1$ in-domain messages in $[d]$.

Analyzer. (Alg. 6) Let C_j be the total number of received messages equal to j. The analyzer outputs the debiased estimate

$$\tilde{f}_j = \frac{C_j}{n} - \frac{k}{2d} \qquad (j \in [d]),$$

which subtracts the known mean noise contribution and does not attempt to identify individual noise messages.

Main result. We defer the proof to App. A.

THEOREM 7 (Low-Communication Histogram Protocol)**.** For any $\varepsilon \in (0,2]$, $\delta \in (0,1)$, and $n > \frac{120d}{\varepsilon^2} \log \frac{8}{\delta}$, let $k = \left\lceil \frac{240d}{\varepsilon^2 n} \log \frac{8}{\delta} \right\rceil$ and $p = \frac{96d}{\varepsilon^2 nk} \log \frac{8}{\delta}$. Then the histogram protocol $\mathcal{P}^{H1} = (\mathcal{R}^{H1}_{b,j,p}, \mathcal{S}, \mathcal{A}^{H1})$ satisfies:

1. **Privacy.** $\mathcal{P}^{H1}$ satisfies (ε, δ)-DP.
2. **Accuracy.** For any $\beta \in (0,1)$, with probability at least $1-\beta$ over the protocol randomness, for each fixed bin $j \in [d]$, $\left|\tilde{f}_j - f_j\right| = O\left(\frac{1}{\varepsilon n}\sqrt{\log \frac{1}{\delta} \cdot \log \frac{1}{\beta}}\right)$. In particular, $\|\tilde{f} - f\|_\infty = O\left(\frac{1}{\varepsilon n}\sqrt{\log \frac{1}{\delta} \cdot \log \frac{d}{\beta}}\right)$
3. **Robustness.** Under any poisoning attack corrupting m users (each limited to at most $k+1$ in-domain messages), the expected estimate of each bin can change by at most $\frac{(k+1)m}{n}$. The expected ℓ_1-deviation of the whole histogram estimate is bounded by $\left\|\mathbb{E}[\tilde{f}] - \mathbb{E}[\tilde{f}^{\mathsf{Adv}}]\right\|_1 \le \frac{2(k+1)m}{n}$.
4. **Communication.** Each user sends at most $k+1$ messages; the expected number is $\frac{k}{2}+1$. Each message carries a bin label in $[d]$ and thus has length $O(\log d)$.

REMARK 2. The lower bound $n > \frac{120d}{\varepsilon^2} \log \frac{8}{\delta}$ confines the protocol to the low-communication regime $k \le 2$; the privacy argument in App. A only requires k and p to be defined as above and establishes the same (ε, δ)-DP guarantee for any $n \in \mathbb{N}$, at the cost of larger k.

5.2 Compressed Histogram for Large Domains

When $d \gg n$, even low-communication protocols may require a larger k to provide privacy. We therefore introduce a compressed histogram protocol that hashes values into a smaller domain $[d_h]$ and performs aggregation in the hashed space, followed by a debiasing reconstruction.

Hash family. We use a pairwise-independent hash family $\{h_{u,v}\}$ mapping $[d]$ to $[d_h]$ such that each $h_{u,v}(x)$ is uniform on $[d_h]$ and $\Pr[h_{u,v}(x) = h_{u,v}(y)] = 1/d_h$ for any $x \neq y$.

Algorithm 7 $\mathcal{R}^{H2}_{b,k,p,\gamma}$: randomizer for compressed histogram

Input: $x \in [d]$, assigned mode flag $b \in \{0,1\}$
Output: multiset $\mathbf{y}$ of triples (u,v,w) with hash seed (u,v) and label $w \in [d_h]$
Sample (u,v) uniformly
Append $(u,v,h_{u,v}(x))$ to $\mathbf{y}$
if $b = 0$ **then**
    $p_b \leftarrow p$
else
    $p_b \leftarrow 1-p$
end if
for $t \leftarrow 1$ **to** k **do**
    if $\mathrm{Ber}(\gamma p_b) = 1$ **then**
        Sample $u,v,w \leftarrow \mathcal{U} \times \mathcal{V} \times [d_h]$ uniformly
        Append (u,v,w) to $\mathbf{y}$
    end if
end for
Return $\mathbf{y}$

Algorithm 8 $\mathcal{A}^{H2}$: analyzer for compressed histogram

Input: shuffled multiset $\mathbf{Y} \subseteq \mathcal{U} \times \mathcal{V} \times [d_h]$
Output: estimate $\tilde{f} \in \mathbb{R}^d$
for $j \in [d]$ **do**
    $C_j \leftarrow \left|\{(u,v,w) \in \mathbf{Y} : h_{u,v}(j) = w\}\right|$
    $\tilde{f}_j \leftarrow \frac{C_j/n - \gamma k/(2d_h) - 1/d_h}{1 - 1/d_h}$
end for
Return $\tilde{f}$

Protocol idea. Each user sends one raw hashed report $(u,v,h_{u,v}(x_i))$. Given mode flag b, let $p_b = p$ if $b = 0$ and $p_b = 1-p$ otherwise. The user then runs k Bernoulli noise trials, each succeeding with probability γp_b (the rate p_b of $\mathcal{P}^{H1}$ scaled by $\gamma \in (0,1]$); on success it samples a fresh (u,v) and $w \leftarrow [d_h]$ and sends (u,v,w). Thus the noise messages are data-independent and uniformly distributed over the compressed message space, with γ controlling the effective noise-message rate. The analyzer reconstructs unbiased estimates by subtracting the known mean noise contribution and correcting for hash collisions in expectation.

Main result. We defer the proof to App. B.

THEOREM 8 (Compressed Histogram Protocol)**.** Fix $\varepsilon \in (0,3]$, $\delta \in (0,1)$ and an integer $d_h \geq 2$. Assume the hash family $\mathcal{H} = \{h_{u,v} : [d] \to [d_h]\}$ is pairwise independent in the following exact sense: for every fixed $x \in [d]$, $h_{u,v}(x)$ is uniform in $[d_h]$ over uniform (u,v), and for every $x \neq y$, $\Pr[h_{u,v}(x) = h_{u,v}(y)] = \frac{1}{d_h}$. Let $k = \left\lceil \frac{108 d_h}{\varepsilon^2 n} \log \frac{8}{\delta} \right\rceil$, $\gamma = \frac{1}{k} \cdot \frac{108 d_h}{\varepsilon^2 n} \log \frac{8}{\delta} \in (0,1]$, and $p \in (0,1/2]$. Then the histogram protocol $\mathcal{P}^{H2} = (\mathcal{R}^{H2}_{b,k,p,\gamma}, \mathcal{S}, \mathcal{A}^{H2})$ satisfies:

1. **Privacy.** $\mathcal{P}^{H2}$ satisfies (ε,δ)-DP.
2. **Accuracy.** For any $\beta \in (0,1)$, with probability at least $1-\beta$ over the protocol randomness, for each fixed item $j \in [d]$, $|\tilde{f}_j - f_j| = O\left(\sqrt{\frac{\log(1/\beta)}{n d_h}} + \frac{1}{\varepsilon n}\sqrt{\log\frac{1}{\delta} \cdot \log\frac{1}{\beta}}\right)$. In particular, $\|\tilde{f} - f\|_\infty = O\left(\sqrt{\frac{\log(d/\beta)}{n d_h}} + \frac{1}{\varepsilon n}\sqrt{\log\frac{1}{\delta} \cdot \log\frac{d}{\beta}}\right)$.
3. **Robustness.** Under any poisoning attack corrupting m users (each still limited to at most $k+1$ in-domain messages), the expected estimate of each item can change by at most $\frac{(k+1)m}{n(1-1/d_h)}$. Moreover, the expected ℓ_1-deviation of the whole estimate satisfies $\left\|\mathbb{E}[\tilde{f}] - \mathbb{E}[\tilde{f}^{\mathsf{Adv}}]\right\|_1 \leq \frac{d(k+1)m}{n(1-1/d_h)}$.
4. **Communication.** Each user sends at most $k+1$ messages; the expected number is $1 + \frac{\gamma k}{2} = 1 + \frac{54 d_h}{\varepsilon^2 n} \log \frac{8}{\delta}$. Each message has length $O(\log d_h)$ up to the seed-length constant.

Whereas $\mathcal{P}^{H1}$ requires $k \propto d$ noise trials, hashing lets $\mathcal{P}^{H2}$ use $k \propto d_h$, decoupling per-user communication from the domain size. The hashed dimension d_h thus tunes an accuracy–communication trade-off, with a larger d_h lowering the hash-collision error ($\propto 1/\sqrt{d_h}$) while raising the expected message count $1 + \gamma k/2$ (linear in d_h).

The compressed histogram protocol $\mathcal{P}^{H2}$ uses our symmetric binomial-sum noise inside the per-message-seed hashing wrapper of Luo et al. [23], where every report carries its own seed and Alg. 8 counts messages whose label matches the query under that message's seed. The large-domain experiment in Sec. 6 therefore compares this noise design against their blanket noise under matched hashing.

6 Experimental Evaluation

We evaluate the proposed histogram protocols on accuracy, communication, and poisoning robustness. The low-communication histogram protocol $\mathcal{P}^{H1}$ is evaluated on the IPUMS *City* dataset [26] ($d = 40$, $n = 155{,}782$), the San Francisco Fire Department *Fire* dataset [16] ($d = 272$, $n = 681{,}174$), and the ACS PUMS *Occupation* dataset [27] ($d = 529$, $n = 123{,}293$). The compressed histogram protocol $\mathcal{P}^{H2}$ is evaluated on the AOL query log [25] with $d = 2^{24}$ and $n = 10^5$. For the three categorical datasets, we fix $\delta = 10^{-6}$, consistent with the common choice $\delta < 1/n$, and vary $\varepsilon \in \{0.25, 0.5, 0.75, 1, 2, 3\}$. For AOL, we set $\delta = 10^{-10}$ and use two hashed sub-domain sizes $d_h \in \{8685, 65\}$, matching LWY's AOL parameter setting [23]. All numbers are means over 100 independent runs.

For $\mathcal{P}^{H1}$, we choose the smallest k for which the joint two-bin numerical condition in Lemma 4 holds for some $p \in (0,1/2]$. This condition certifies (ε,δ)-DP at each evaluated ε, including $\varepsilon = 3$, which lies beyond the $\varepsilon \leq 2$ range of Theorem 7. In the AOL experiment, we calibrate both protocols with LWY's numerical search rule [23], so that the comparison is made under a single parameter-selection criterion. The resulting message counts are far smaller than implied by the conservative closed-form parameters of Theorem 8.

We measure accuracy by count-level MAE over histogram bins, reported in absolute counts rather than normalized frequencies. Communication is the average number of messages sent by one honest user, counting both data and noise messages. For robustness, we report the empirical poisoning influence, namely the increase in normalized ℓ_1 histogram error after replacing a uniformly random m/n fraction of users by corrupted users that submit the maximum number of syntactically valid messages the protocol accepts, with $m/n \in \{0.1\%, 1\%, 5\%, 10\%, 20\%, 30\%\}$. This is a lower bound on the worst-case influence $\mathrm{Infl}(m)$.

Small-domain evaluation. Tab. 2 reports accuracy (MAE) and communication (messages per user) of $\mathcal{P}^{H1}$ on the three categorical datasets across privacy levels. It shows that $\mathcal{P}^{H1}$ substantially improves accuracy over CSUZZ [10], BC [2], and CZ [11]. Across all three datasets and all reported privacy levels, the MAE reduction is between $4.2\times$ and $13.2\times$. For example, at $\varepsilon = 1$ on City, $\mathcal{P}^{H1}$ has MAE 5.4, compared with 46, 44, and 25 for CSUZZ, BC, and CZ, respectively. The same pattern holds on Fire and Occupation, even though their domain sizes differ by more than an order of magnitude. GKMP [18] attains the smallest MAE but only with a much larger message budget, ranging from thousands to millions of messages per user, placing it at a different accuracy–communication point.

The message counts should also be read together with the per-message format. CSUZZ and BC use scalar messages but require d-scale communication, while CZ sends only two messages but each message is a d-bit vector. The corresponding bit-level cost is reported in the full version of this paper. Thus, among protocols with short scalar messages and small expected message counts, the closest comparison is LWY [23]. At $\varepsilon = 0.25$, $\mathcal{P}^{H1}$ achieves MAE essentially equal to LWY's across the three datasets. As ε increases, $\mathcal{P}^{H1}$ becomes consistently more accurate. At $\varepsilon = 1$, it reduces MAE relative to LWY by 10.0% on City, 10.2% on Fire, and 8.3% on Occupation. At $\varepsilon = 2$, the reduction is about 29–32%, and at $\varepsilon = 3$, it reaches about 48% across the three datasets. This widening advantage at larger ε is consistent with the different aggregate noise shapes: LWY's centered blanket noise becomes right-

Table 2: Empirical comparison of accuracy and communication for representative shuffle-model protocols across datasets and privacy levels ($\delta = 10^{-6}$ and $\varepsilon \in \{0.25, 0.5, 0.75, 1, 2, 3\}$). Larger ε accentuates the MAE advantage of Ours over LWY (e.g., $\sim$48% reduction at ε=3 on all three datasets), reflecting that our symmetric binomial-sum noise avoids the right skew that LWY's blanket noise develops at small per-bin means.

	MAE						Msgs/user					
Protocol	0.25	0.5	0.75	1	2	3	0.25	0.5	0.75	1	2	3
Dataset: City with d=40 and $n\approx$156k												
CSUZZ [10]	280	101	62	46	22	14	40	40	40	40	40	40
BC [2]	148	85	58	44	22	15	29	38	40	40	41	41
GKMP [18]	7.9	4.0	2.6	1.9	0.9	0.5	416k	107k	49k	29k	8.0k	4.1k
CZ [11]	99	48	32	25	15	12	2	2	2	2	2	2
LWY [23]	21.6	11.1	7.7	6.0	3.8	3.1	1.19	1.05	1.02	1.01	1.01	1.00
Ours	22.0	10.9	7.3	5.4	2.6	1.6	1.5	1.5	1.5	1.5	1.5	1.5
Dataset: Fire with d=272 and $n\approx$681k												
CSUZZ [10]	194	89	59	43	22	14	272	272	272	272	272	272
BC [2]	169	87	58	44	22	15	254	268	271	272	273	273
GKMP [18]	8.0	4.0	2.6	1.9	0.9	0.5	647k	166k	76k	45k	12k	6.3k
CZ [11]	93	46	32	24	15	12	2	2	2	2	2	2
LWY [23]	21.6	11.0	7.6	5.9	3.7	3.1	1.29	1.08	1.04	1.02	1.01	1.01
Ours	21.4	10.8	7.1	5.3	2.6	1.6	2.0	1.5	1.5	1.5	1.5	1.5
Dataset: Occupation with d=529 and $n\approx$123k												
CSUZZ [10]	285	102	62	45	22	14	529	529	529	529	529	529
BC [2]	136	83	57	43	22	15	321	478	507	517	527	529
GKMP [18]	8.0	4.0	2.6	1.9	0.9	0.5	6.9M	1.8M	818k	480k	134k	68k
CZ [11]	102	48	32	25	15	12	2	2	2	2	2	2
LWY [23]	21.5	11.1	7.7	6.0	3.7	3.1	4.15	1.83	1.39	1.24	1.09	1.06
Ours	21.6	10.8	7.2	5.5	2.6	1.6	7.5	3.0	2.0	1.5	1.5	1.5

Table 3: Large-domain accuracy and communication comparison on AOL, with the same experimental setup as LWY [23] (d=2^{24}, n=10^5, δ=10^{-10}, $d_h \in \{8685, 65\}$). Ours matches LWY on both metrics across $\varepsilon \in \{0.25, 0.5, 0.75, 1, 2, 3\}$.

	MAE						Msgs/user					
Protocol	0.25	0.5	0.75	1	2	3	0.25	0.5	0.75	1	2	3
Hashed sub-domain d_h=$n/\log n\approx$8685												
LWY [23]	28.9	15.0	10.5	8.4	5.6	4.9	114	30.7	15.0	9.5	4.2	3.3
Ours	28.9	15.0	10.5	8.4	5.6	4.9	114	30.7	15.0	9.5	4.2	3.3
Hashed sub-domain d_h=$n/\log^3 n\approx$65												
LWY [23]	42.9	34.9	33.2	32.5	31.9	31.8	1.85	1.22	1.11	1.06	1.02	1.02
Ours	42.9	34.9	33.2	32.6	31.9	31.8	1.85	1.22	1.11	1.06	1.02	1.02

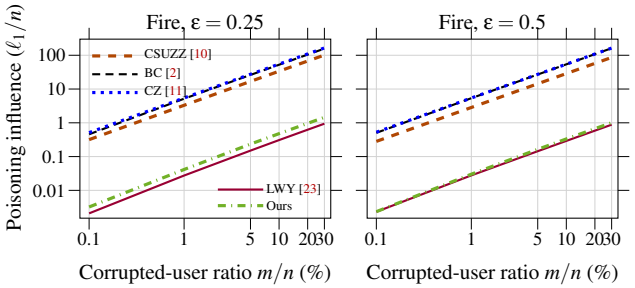

Figure 2: Poisoning influence comparison on the Fire dataset at $\varepsilon \in \{0.25, 0.5\}$. Ours and LWY remain orders of magnitude below CSUZZ, BC, and CZ across the full m/n range. The full figure for City, Fire, and Occupation at $\varepsilon \in \{0.25, 0.5, 0.75, 1\}$ appears in the full version of this paper.

skewed when the per-bin mean is small, whereas $\mathcal{P}^{H1}$ uses symmetric binomial-sum noise.

In terms of communication, $\mathcal{P}^{H1}$ remains in the same low-message range as LWY in the evaluated settings. It uses 1.5 messages per user on all three datasets once $\varepsilon \geq 1$; the strongest-privacy regime requires a few additional messages (at most 7.5 per user) to satisfy the DP target. Overall, the small-domain experiments show a favorable trade-off. Our $\mathcal{P}^{H1}$ keeps communication low, improves substantially over CSUZZ/BC/CZ, and improves over LWY at all but $\varepsilon = 0.25$, where the two achieve essentially equal MAE.

Large-domain evaluation. We use AOL ($d = 2^{24}$) to test whether the compressed protocol $\mathcal{P}^{H2}$ preserves LWY's large-domain accuracy–communication trade-off under the same settings. The larger d_h reduces hash-collision error at higher communication, while the smaller favors near-single-message communication. Running both protocols at the same d_h isolates the noise-design contribution by holding the hash-collision rate $1/d_h$ identical across the two protocols. As shown in Tab. 3, $\mathcal{P}^{H2}$ matches LWY's MAE and messages per user under this matched setup.

Poisoning-influence evaluation. Fig. 2 reports representative poisoning-influence curves on the Fire dataset under varying corrupted-user ratios m/n from 0.1% to 30%, and the full version of this paper gives the full set of results for all three datasets and $\varepsilon \in \{0.25, 0.5, 0.75, 1\}$. Across these settings, the curves grow approximately linearly in m/n, as predicted by the per-user message-limit analysis. On poisoning influence, our $\mathcal{P}^{H1}$ stays close to LWY [23], the strongest baseline, and is consistently far below CSUZZ [10], BC [2], and CZ [11]. The reason is structural: CSUZZ, BC, and CZ let a corrupted user push every coordinate (via d-bit reports or d-dim vectors), so their poisoning grows with d; LWY and $\mathcal{P}^{H1}$ cap each user at a few scalar messages, bounding the coordinates corruption can touch. For example, at $m/n = 10\%$ and $\varepsilon = 0.25$ on Fire, $\mathcal{P}^{H1}$ reaches 0.48 vs CSUZZ 33.72, BC 52.39, CZ 54.80; LWY is comparable at 0.31, and the gap between $\mathcal{P}^{H1}$ and LWY narrows as ε increases. We omit GKMP [18] from Fig. 2 because it has no deterministic per-user message limit, so a corrupted user could repeat in-domain messages and make the worst-case poisoning influence unbounded by repetition.

Taken together, these results establish that our protocols achieve the best accuracy-communication-robustness trade-off, delivering up to nearly 2$\times$ lower small-domain MAE than

the closest baseline [23], with comparable per-user communication and poisoning influence.

7 Discussion

7.1 Practical Considerations of Preprocessing

In our protocols, preprocessing is used to instantiate the data-independent auxiliary-input distribution required by Def. 3. We discuss its impact on latency, coordination state, and user dropout in the three subsections that follow.

Latency. Because the auxiliary inputs are independent of user data, they can be generated and delivered before the analytics window starts, or batched with other round-configuration material. Thus, in scheduled collection workflows, preprocessing need not lie on the critical path of user reporting; in interactive settings, it adds one setup phase before the usual collection round and incurs no additional per-round latency.

Coordination state. The additional state is temporary and small. During preprocessing, the shuffler holds the auxiliary-input multiset and the random assignment used for delivery; this state can be erased once delivery completes. Each user stores only its own auxiliary input until reporting finishes. The analyzer needs only the public round parameters and the prescribed auxiliary-input distribution, but must not learn the per-user assignment.

User dropout. A dropped user contributes no report, and the remaining reports are processed normally. This applies whether the user drops out during preprocessing, between preprocessing and reporting, or during reporting itself. As with other shuffle protocols whose guarantees depend on the number of submitted reports, our privacy and accuracy guarantees should be instantiated for the effective reporting population. In practice, a deployment may choose parameters for a conservative lower bound on the number of reporters, or restart the round if participation falls below the required threshold.

7.2 Deployment Compatibility

The main deployment question is whether an implementation provides a setup channel for distributing data-independent auxiliary inputs before users report. Otherwise, the flag-free split-trial variant of Remark 1 remains available under the standard shuffle model, with higher communication and a looser poisoning-influence bound.

Prochlo-style deployments. A Prochlo-style Encode-Shuffle-Analyze architecture [6] is compatible with our design if auxiliary inputs are distributed as configuration material before the Encode step. The analyzer may specify the data-independent auxiliary-input multiset required by Def. 3, but must not learn which user receives which element; preprocessing then amounts to a limited extension of the existing setup channel used to configure each collection round.

Mix-net-based deployments. A canonical mix-net [8] is primarily one-way, with senders submitting messages, the mix-net anonymizes them, and recipients obtain the outputs. It does not by itself expose a shuffler-to-sender delivery path. A strictly one-way mix-net deployment would therefore need a separate setup channel to realize our preprocessing step; otherwise, the flag-free split-trial variant of Remark 1 remains the natural fallback at the trade-offs noted above.

TEE-based deployments. A shuffler implemented inside a trusted execution environment (TEE) [12] can realize the preprocessing step as enclave-internal logic. The enclave obtains or samples the data-independent auxiliary-input multiset during setup and delivers the assigned auxiliary input to each user over a channel established via remote attestation. This requires no separate setup channel, subject to the usual TEE-deployment assumptions (enclave integrity, attestation soundness, and a trusted code base).

8 Conclusion

We proposed shuffle protocols that achieve (ε, δ)-DP and strong accuracy while providing explicit poisoning-influence bounds under the per-user message-limit premise. Our design centers on a *symmetric binomial-sum noise* construction realized via a shuffled assignment of data-independent auxiliary inputs, enabling a simple constant-shift estimator and explicit worst-case bias analysis under anonymity. We provide formal privacy, accuracy, and robustness guarantees for binary estimation and histograms (including a large-domain variant) and validate the resulting accuracy–communication–robustness trade-offs on real datasets. More broadly, our results highlight a protocol-design principle for the shuffle model: beyond calibrating noise for DP, structurally limiting how much each user can report is key to provably bounding worst-case estimator bias under anonymity. From this perspective, symmetric binomial-sum noise provides a reusable design building block. Its known mean shift supports low-variance estimation without multiplicative bias amplification, and it combines with domain-reduction techniques such as hashing while retaining explicit influence bounds under the standard per-user message limit.

Ethical Considerations

Scope and intended use. This work develops and evaluates shuffle-model differential privacy protocols for frequency estimation and histogram construction, with emphasis on mitigating the impact of poisoning by corrupted users. Our defensive goal is to improve the reliability of privacy-preserving aggregate analytics under realistic adversarial behavior.

Stakeholders, benefits, and potential harms. This work affects several stakeholder groups, directly or indirectly: (i) *data subjects* represented in the public evaluation datasets,

and more broadly populations whose aggregate statistics may be released using shuffle-model protocols; (ii) *deploying organizations* that operate shuffler and analyzer components or use such protocols for aggregate analytics; (iii) *downstream consumers* of the released aggregate statistics, including the communities those statistics describe; (iv) the broader *differential-privacy and shuffle-model research community*. The intended benefit across these groups is more reliable privacy-preserving aggregate analytics: stronger robustness against tampering with aggregate releases, and reusable design primitives for future shuffle-model protocols. Potential harms include dual-use concerns arising from the poisoning analysis, over-interpretation of our worst-case bounds as deployment guarantees outside the stated assumptions, deployment without the required per-user message-bound enforcement, and downstream reliance on biased or noisy aggregates.

Datasets, human subjects, and privacy. We do not recruit participants, interact with human subjects, or collect new personal data. The small-domain evaluation uses three publicly available categorical datasets, IPUMS City, SFFD Fire, and ACS PUMS Occupation (all 2023; $d = 40,272,529$), each contributing one categorical attribute per user, with no personally identifiable information (PII) for our purposes and used solely for research and benchmarking under their stated terms. The large-domain evaluation uses the AOL query log [25], the standard large-domain benchmark in this line of work [23]. Although the 2006 release was withdrawn after users were re-identified from raw query text and per-user metadata, our evaluation reads only a histogram of fixed-length query prefixes over $d = 2^{24}$ keys (no raw queries, identifiers, timestamps, or click-throughs), so it retains none of the fields that enabled re-identification. Matching LWY's dataset keeps the comparison in Tab. 3 valid.

Data minimization and artifact release. To minimize privacy and misuse risks, our released artifacts contain no record-level data, only the aggregated input histograms used by our evaluation (categorical frequency vectors for the three small-domain datasets and the 3-character query-prefix histogram described above for AOL); preprocessing scripts that read record-level inputs are not redistributed.

Dual-use considerations: poisoning analysis. Because the paper studies poisoning strategies and quantifies their influence, there is a potential dual-use concern that such analyses could inform attempts to bias aggregate statistics. We mitigate this risk by focusing on mitigation mechanisms and bounded-influence designs, and by scoping released code to research reproduction. In particular, our poisoning code is provided only as a simulation module for controlled benchmarking of robustness claims; it is not intended as tooling for attacking deployed systems. On balance, we judge publication to be net beneficial, since poisoning can undermine private aggregate analytics even when individual privacy is preserved, and openly articulating the assumptions, limitations, and bounded-influence mechanisms enables scrutiny, reproducibility, and clearer guidance for future shuffle-model protocol design.

Responsible experimentation. All experiments are conducted offline within our evaluation framework. We do not probe, attack, or measure any real-world systems without authorization, and we do not disclose vulnerabilities in specific deployed services.

Open Science

Artifact availability. We release a permanent, versioned artifact on Zenodo with a citable DOI:

https://doi.org/10.5281/zenodo.20405419

The artifact contains Python implementations of our proposed shuffle protocols and the baselines evaluated in the paper, including LWY, CSUZZ, BC, CZ, and GKMP. It also includes a top-level `README.md` describing the code structure, configuration parameters, and commands needed to reproduce the reported results. A full version of this paper, with all proofs, is available at https://eprint.iacr.org/2026/109957.

Reproducibility scope. The released scripts reproduce the main evaluation metrics in the paper, including error, messages per user, and robustness under poisoning simulations. Because the protocols and poisoning experiments involve randomized procedures, repeated runs may exhibit small statistical variation.

Acknowledgments

We thank our shepherd and the reviewers for their constructive comments and guidance throughout the revision. This work is supported by the National Cryptologic Science Fund of China under Grant 2025NCSF01010, and the Key Program of the National Natural Science Foundation of China under Grants U25B2028 and 62432012.

References

[1] Apple Differential Privacy Team. Learning with privacy at scale. Apple Machine Learning Research, 2017. URL: https://machinelearning.apple.com/research/learning-with-privacy-at-scale.

[2] Victor Balcer and Albert Cheu. Separating local & shuffled differential privacy via histograms. In *Information-Theoretic Cryptography (ITC)*, 2020.

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

[4] Borja Balle, James Bell, Adrià Gascón, and Kobbi Nissim. The privacy blanket of the shuffle model. In *CRYPTO*, pages 638–667, 2019.

[5] Borja Balle, James Bell, Adria Gascón, and Kobbi Nissim. Private summation in the multi-message shuffle model. In *ACM CCS*, pages 657–676, 2020.

[6] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Operating Systems Principles*, pages 441–459, 2017.

[7] Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Data poisoning attacks to local differential privacy protocols. In *USENIX Security*, pages 947–964, 2021.

[8] David L. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 24(2):84–90, 1981.

[9] Albert Cheu, Adam Smith, and Jonathan Ullman. Manipulation attacks in local differential privacy. In *IEEE SP*, pages 883–900, 2021.

[10] Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *EUROCRYPT*, pages 375–403, 2019.

[11] Albert Cheu and Maxim Zhilyaev. Differentially private histograms in the shuffle model from fake users. In *IEEE SP*, pages 440–457, 2022.

[12] Victor Costan and Srinivas Devadas. Intel SGX explained. Cryptology ePrint Archive, Paper 2016/086, 2016. URL: https://eprint.iacr.org/2016/086.

[13] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

[14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[15] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *ACM CCS*, pages 1054–1067, 2014.

[16] Fire Department Calls for Service. San francisco fire department calls for service [dataset]. https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3/about_data, 2023.

[17] Badih Ghazi, Noah Golowich, Ravi Kumar, Rasmus Pagh, and Ameya Velingker. On the power of multiple anonymous messages: Frequency estimation and selection in the shuffle model of differential privacy. In *EUROCRYPT*, pages 463–488, 2021.

[18] Badih Ghazi, Ravi Kumar, Pasin Manurangsi, and Rasmus Pagh. Private counting from anonymous messages: Near-optimal accuracy with vanishing communication overhead. In *Proceedings of Machine Learning Research (PMLR)*, pages 3505–3514, 2020.

[19] Badih Ghazi, Ravi Kumar, Pasin Manurangsi, Rasmus Pagh, and Amer Sinha. Differentially private aggregation in the shuffle model: Almost central accuracy in almost a single message. In *Proceedings of Machine Learning Research (PMLR)*, pages 3692–3701, 2021.

[20] Michael B Hawes. Implementing differential privacy: Seven lessons from the 2020 united states census. *Harvard Data Science Review*, 2(2), 2020.

[21] Yuval Ishai, Eyal Kushilevitz, Rafail Ostrovsky, and Amit Sahai. Cryptography from anonymity. In *Foundations of Computer Science (FOCS)*, pages 239–248, 2006.

[22] Xiaoguang Li, Ninghui Li, Wenhai Sun, Neil Zhenqiang Gong, and Hui Li. Fine-grained poisoning attack to local differential privacy protocols for mean and variance estimation. In *USENIX Security*, pages 1739–1756, 2023.

[23] Qiyao Luo, Yilei Wang, and Ke Yi. Frequency estimation in the shuffle model with almost a single message. In *ACM CCS*, pages 2219–2232, 2022.

[24] Takao Murakami, Yuichi Sei, and Reo Eriguchi. Augmented shuffle protocols for accurate and robust frequency estimation under differential privacy. In *IEEE SP*, pages 3892–3911, 2025.

[25] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems (InfoScale)*, pages 1–7, 2006.

[26] Sarah Flood Steven Ruggles et al. Ipums usa: Version 14.0 [dataset]. https://www.ipums.org/projects/ipums-usa/d010.V14.0, 2023.

[27] U.S. Census Bureau. American community survey public use microdata sample (pums) [dataset]. https://www.census.gov/programs-surveys/acs/microdata.html, 2023.

[28] Yongji Wu, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Poisoning attacks to local differential privacy protocols for key-value data. In *USENIX Security*, pages 519–536, 2022.

A Proof of Theorem 7

We give the proof for the low-communication histogram protocol. Complete algebraic details and routine concentration calculations appear in the full version of this paper. For clarity, we write the proof for the exactly balanced case where n is divisible by $2d$; rounding changes only constants.

Noise characterization. Fix a bin $j \in [d]$ and let $T_j := \sum_{i=1}^{n} \mathbf{1}[x_i = j]$ be the true count in that bin. The preprocessing assigns exactly $n/(2d)$ users with mode 0 and $n/(2d)$ users with mode 1 to bin j. Since each assigned user performs k noise trials, the two mode groups contribute

$$N_j^{(0)} \sim \mathrm{Bin}\left(\frac{nk}{2d}, p\right), \qquad N_j^{(1)} \sim \mathrm{Bin}\left(\frac{nk}{2d}, 1-p\right),$$

independently. Thus, with $M := nk/(2d)$,

$$C_j = T_j + N_j, \qquad N_j := N_j^{(0)} + N_j^{(1)}, \qquad \mathbb{E}[N_j] = M.$$

The analyzer therefore satisfies

$$\tilde{f}_j = \frac{C_j}{n} - \frac{k}{2d} = \frac{T_j}{n} + \frac{N_j - M}{n},$$

so the estimator is unbiased.

Privacy. A neighboring histogram change moves one raw report from an old bin to a new bin. Hence only two bins can have data-dependent counts changed; all other coordinates contain the same raw contribution and data-independent noise. For a fixed affected bin, the released count is a post-processing of the two-count release

$$(T_0 + X, T_1 + Y), \qquad X \sim \mathrm{Bin}(M, p), \quad Y \sim \mathrm{Bin}(M, 1-p),$$

where $T_b := |\{i : x_i = j, b_i = b\}|$ counts users with raw value j and mode b, irrespective of the assigned noise bin. Between two neighboring inputs, exactly one of T_0, T_1 changes by 1, with mode 0 vs 1 each having probability $1/2$ by the data-independent uniform mode assignment, providing the $1/2$-symmetric structure required by Theorem 5. Instantiating Theorem 5, whose two-directional DP is established via Lemma 3, per bin with privacy parameters $(\varepsilon/2, \delta/2)$ and effective sample size $n_{\mathrm{bin}} = 2M = nk/d$ gives the fixed value

$$p = \frac{24}{(\varepsilon/2)^2 n_{\mathrm{bin}}} \log \frac{4}{\delta/2} = \frac{96d}{\varepsilon^2 nk} \log \frac{8}{\delta},$$

provided $\varepsilon \le 2$ (so $\varepsilon/2 \le 1$) and $p \le 2/5$ (Lemma 3's hypothesis), the latter ensured by

$$k \ge \frac{240d}{\varepsilon^2 n} \log \frac{8}{\delta}$$

(under which $p = 96/240 = 2/5$ at the boundary). Thus the stated choice of k and the fixed p above gives $(\varepsilon/2, \delta/2)$-DP for each affected bin. Basic composition over the at most two affected bins gives (ε, δ)-DP for the full histogram release, and the analyzer output is post-processing. Theorem 8's direct Chernoff proof in App. B admits $\varepsilon \le 3$ at the cost of looser per-bin constants; this reduction-based proof prioritizes tightness in the small-domain regime.

Accuracy. The deviation of a fixed bin is $(N_j - M)/n$. Since $N_j - M$ is a sum of $2M$ independent, centered, $[-1, 1]$-bounded variables and

$$\mathrm{Var}(N_j) = 2Mp(1-p),$$

Bernstein's inequality gives, with probability at least $1 - \beta$,

$$|\tilde{f}_j - f_j| \le \frac{1}{3n} \log \frac{2}{\beta} + \frac{1}{n} \sqrt{\left(\frac{1}{3} \log \frac{2}{\beta}\right)^2 + 4Mp(1-p) \log \frac{2}{\beta}}.$$

For the fixed p from Theorem 5, we have

$$Mp(1-p) \le Mp = \frac{48}{\varepsilon^2} \log \frac{8}{\delta}.$$

Substitution yields the fixed-bin bound

$$|\tilde{f}_j - f_j| = O\left(\frac{1}{\varepsilon n} \sqrt{\log \frac{1}{\delta} \log \frac{1}{\beta}}\right).$$

Applying the same bound with failure probability β/d and union-bounding over $j \in [d]$ gives the ℓ_∞ guarantee.

Robustness and communication. Under the per-user message limit, each corrupted user can contribute at most $k+1$ in-domain messages. For a fixed bin, the per-bin influence counts only the worst injection into that bin. Across m corrupted users C_j can change by at most $(k+1)m$, and the affine estimator scales counts by $1/n$, giving influence at most $(k+1)m/n$ for that bin. For the full-histogram ℓ_1 deviation, both deletions and insertions count. Replacing one honest multiset by an adversarial multiset of size at most $k+1$ contributes at most $k+1$ deletions and $k+1$ insertions, so the ℓ_1 change in the count vector is at most $2(k+1)$ per corrupted user. Hence

$$\left\| \mathbb{E}[\tilde{f}] - \mathbb{E}[\tilde{f}^{\mathsf{Adv}}] \right\|_1 \le \frac{2(k+1)m}{n}.$$

Each honest user always sends one raw message and has k noise trials. Since the balanced mode assignment makes the average success probability $(p + (1-p))/2 = 1/2$, the expected number of messages is $1 + k/2$, and each message carries a label in $[d]$, requiring $O(\log d)$ bits.

LEMMA 4 (Joint two-bin numerical (ε,δ)-DP check)**.** Let π be the pmf of $Z = \mathrm{Bin}(M,p) + \mathrm{Bin}(M,1-p)$ and define the boundary-extended log-ratio

$$g(t) := \log \frac{\pi(t-1)}{\pi(t)}, \qquad g(0) := -\infty, \quad g(2M+1) := +\infty.$$

If independent $Z_A, Z_B \sim \pi$ satisfy

$$\Pr[g(1+Z_A) - g(Z_B) > \varepsilon] \le \delta,$$

then the two affected bins satisfy the required joint (ε,δ) privacy condition.

Proof. A neighboring change differs at one user and affects two bins, with released counts c_A, c_B. The joint two-bin privacy loss equals $\log[\pi(c_A-1)\pi(c_B)/(\pi(c_A)\pi(c_B-1))] = g(c_A) - g(c_B)$, which under the changed dataset realizes $g(1+Z_A) - g(Z_B)$ with $Z_A, Z_B \sim \pi$ independent. The hypothesis bounds $\Pr[g(1+Z_A) - g(Z_B) > \varepsilon] \le \delta$, yielding (ε,δ)-DP via the standard tail-probability sufficient condition [14]; the reverse direction is symmetric. □

The closed-form choice above is a sufficient analytic guarantee; Lemma 4 is used only for data-independent parameter selection in experiments.

B Proof of Theorem 8

We give the proof for the compressed histogram protocol. Complete multinomial algebra and routine tail-bound details appear in the full version of this paper. Let $p_c := 1/d_h$ and $\mu := n\gamma k/(2d_h)$. For a message $y = (u,v,t)$, write

$$I_y(j) := \mathbf{1}[t = h_{u,v}(j)], \qquad C_j := \sum_{y \in \mathbf{Y}} I_y(j).$$

The analyzer computes

$$\tilde{f}_j = \frac{C_j/n - \gamma k/(2d_h) - p_c}{1-p_c}.$$

Throughout, we use the exact collision assumption of Theorem 8, so that $p_c = 1/d_h$. (For a concrete family with $p_{\mathrm{col}} \le 1/d_h$ rather than equality, the estimator and proof apply with p_{col} in place of p_c.)

Privacy. Let $\Omega := \mathcal{U} \times \mathcal{V} \times [d_h]$ and

$$S_x := \{(u,v,t) \in \Omega : t = h_{u,v}(x)\}.$$

A raw report for value x is uniform over S_x, while every successful noise message is uniform over Ω and independent of the dataset. For neighboring datasets differing only in one user's value $x \to x'$, the unchanged users' raw reports are identically distributed in both executions, and removing them only makes the privacy analysis harder. It remains to analyze a balls-into-bins mechanism with input set S_x or $S_{x'}$.

For a multiset $W = (w_m)_{m\in\Omega}$, the noise-only probability depends on W only through its total size and multinomial allocation. Therefore, for the mechanism including one raw report,

$$\frac{\Pr[M(S_x) = W]}{\Pr[M(S_{x'}) = W]} = \frac{\sum_{m\in S_x} w_m}{\sum_{m\in S_{x'}} w_m}.$$

When $W \sim M(S_x)$, the changed raw report contributes one message to S_x. Let Z_x and $Z_{x'}$ be the noise-message counts falling in S_x and $S_{x'}$. The bad likelihood-ratio event is contained in

$$\{1 + Z_x \ge e^{\varepsilon} Z_{x'}\},$$

attained at the worst case $S_x \cap S_{x'} = \emptyset$; overlap only reduces this event. For every fixed item, the marginal noise count has distribution

$$Z = \mathrm{Bin}\left(\frac{nk}{2}, \frac{\gamma p}{d_h}\right) + \mathrm{Bin}\left(\frac{nk}{2}, \frac{\gamma(1-p)}{d_h}\right), \qquad \mathbb{E}[Z] = \mu.$$

Using only the marginal Chernoff bounds on Z_x and $Z_{x'}$ with a union bound (no joint independence needed), the bad event has probability at most δ when

$$\mu = \frac{n\gamma k}{2d_h} = \frac{54}{\varepsilon^2}\log\frac{8}{\delta}$$

and $0 < \varepsilon \le 3$ (the upper bound $\varepsilon \le 3$ comes from $1 - \varepsilon/4 > e^{-\varepsilon/2}$ used to combine the two Chernoff events; the difference is concave with value 0 at $\varepsilon = 0$ and $1/4 - e^{-3/2} > 0$ at $\varepsilon = 3$, so the inequality holds on $(0,3]$). The reverse direction is symmetric in x and x', so $\mathcal{P}^{H2}$ satisfies (ε,δ)-DP.

Accuracy. Fix $j \in [d]$ and let $r_j := nf_j$. Decompose C_j into true appearances, hash collisions from non-j raw reports, and noise matches:

$$C_j = r_j + W_j + U_j.$$

Here $W_j \sim \mathrm{Bin}(n - r_j, p_c)$, and

$$U_j = U_j^{(0)} + U_j^{(1)}, \qquad U_j^{(0)} \sim \mathrm{Bin}\left(\frac{nk}{2}, \frac{\gamma p}{d_h}\right),$$
$$U_j^{(1)} \sim \mathrm{Bin}\left(\frac{nk}{2}, \frac{\gamma(1-p)}{d_h}\right).$$

Thus $\mathbb{E}[U_j] = \mu$ and

$$\mathbb{E}[C_j] = r_j + (n - r_j)p_c + \mu.$$

Plugging this expression into the analyzer formula gives $\mathbb{E}[\tilde{f}_j] = f_j$.

Since W_j and U_j are sums of independent indicators, Bernstein's inequality gives, with probability at least $1-\beta$,

$$|W_j - \mathbb{E}W_j| = O\left(\sqrt{np_c \log\frac{1}{\beta}} + \log\frac{1}{\beta}\right),$$

and

$$|U_j - \mathbb{E}U_j| = O\left(\sqrt{\mu \log \frac{1}{\beta}} + \log \frac{1}{\beta}\right).$$

Dividing by $n(1-p_c)$ and using $p_c = 1/d_h$ and $\mu = O(\varepsilon^{-2}\log(1/\delta))$ gives the three-term bound

$$|\tilde{f}_j - f_j| = O\left(\sqrt{\frac{\log(1/\beta)}{nd_h}} + \frac{1}{\varepsilon n}\sqrt{\log\frac{1}{\delta}\log\frac{1}{\beta}} + \frac{\log(1/\beta)}{n}\right).$$

Theorem 8's statement displays the first two terms for readability; the additive $O(\log(1/\beta)/n)$ term is dominated by the others in the typical small-ε, moderate-δ regime, though not in general. The ℓ_∞ statement follows by applying the fixed-item bound with failure probability β/d and union-bounding over all d items.

Robustness and communication. A corrupted user can contribute at most $k+1$ accepted in-domain messages. For a fixed target item j, each crafted message can be made to match j, so C_j changes by at most $(k+1)m$. The affine estimator scales this count by $1/(n(1-p_c))$, giving fixed-item influence at most

$$\frac{(k+1)m}{n(1-1/d_h)}.$$

For the full vector, summing the fixed-item bound over all d items via the trivial $\ell_1 \le d \cdot \ell_\infty$ inequality gives

$$\left\|\mathbb{E}[\tilde{f}] - \mathbb{E}[\tilde{f}^{\mathsf{Adv}}]\right\|_1 \le \frac{d(k+1)m}{n(1-1/d_h)}.$$

This bound is attained (up to lower-order honest-contribution savings) by the degenerate-seed attack $(u,v) = (0,v_0)$, $t = v_0 \bmod d_h$, under which every adversarial message has $I_y(j) = 1$ for all $j \in [d]$. Each user sends one raw hashed report and then performs k noise trials with success probability γp_b, where $p_b \in \{p, 1-p\}$. The balanced mode assignment gives average success probability $\gamma/2$, so the expected number of messages per user is

$$1 + \frac{\gamma k}{2} = 1 + \frac{54\,d_h}{\varepsilon^2 n}\log\frac{8}{\delta}.$$

Each message contains a hash seed and a label in $[d_h]$, i.e., length $O(\lambda_{\mathcal{H}} + \log d_h)$ bits, where $\lambda_{\mathcal{H}}$ is the seed length of the hash family.

Numerical calibration. The closed-form rate above is sufficient but conservative. For LWY-comparable experiments [23], we use their binomial proxy $X \sim \mathrm{Bin}(n, \mu/n)$ for parameter selection. This proxy matches our true noise distribution Z above only in mean μ and is not a full exact DP verifier for the histogram mechanism.

C Anonymous Message Control

Goal. All poisoning-influence bounds in this paper assume a per-user acceptance limit: in one collection round, at most $k+1$ messages from any admitted participant are accepted by the analyzer. Here $k = 1$ for the binary protocol and $k = \left\lceil \frac{240d}{\varepsilon^2 n}\log\frac{8}{\delta}\right\rceil$ (resp. $\left\lceil \frac{108 d_h}{\varepsilon^2 n}\log\frac{8}{\delta}\right\rceil$) for the low-communication (resp. compressed) histogram protocol. This condition is necessary in the shuffle model. If a corrupted participant can cause arbitrarily many syntactically valid in-domain messages to be accepted, then the worst-case estimator bias is unbounded by repetition. The mechanism below shows one way to realize this premise without requiring the analyzer to identify senders from the shuffled transcript.

Token mechanism. Fix a round identifier eid and a per-user limit $k+1$. The analyzer samples a signature key pair $(\mathsf{sk},\mathsf{pk}) \leftarrow \mathsf{KeyGen}(1^\lambda)$ and creates $n(k+1)$ fresh one-time tokens $\tau_1,\ldots,\tau_{n(k+1)}$. For each token it computes

$$\sigma_i \leftarrow \mathsf{Sign}(\mathsf{sk}, \mathsf{eid} \parallel \tau_i).$$

The analyzer sends the multiset $\{(\tau_i,\sigma_i)\}_{i=1}^{n(k+1)}$ to the shuffler. The shuffler uniformly permutes this multiset and delivers exactly $k+1$ token pairs to each admitted participant. Tokens contain no explicit user identifier; the shuffler knows the delivery assignment but, by the standard non-collusion assumption, does not share it with the analyzer.

Each reported protocol message is submitted as a tuple (m,τ,σ) using one previously unused token pair. After shuffling, the analyzer accepts m only if

$$\mathsf{Vfy}(\mathsf{pk}, \mathsf{eid} \parallel \tau, \sigma) = 1 \quad \text{and} \quad \tau \notin \mathsf{Spent}_{\mathsf{eid}},$$

where $\mathsf{Spent}_{\mathsf{eid}}$ is the set of tokens already accepted in the current round. If the checks pass, the analyzer inserts τ into $\mathsf{Spent}_{\mathsf{eid}}$; otherwise the tuple is discarded.

Guarantee and scope. Assuming signature unforgeability and correct replay checking, each token authorizes at most one accepted message. Therefore, an adversary corrupting m admitted participants obtains at most $(k+1)m$ accepted messages in the round, matching the robustness premise. The mechanism does not address admission control or Sybil resistance; multiple admitted identities scale the bound linearly. It also assumes integrity of token delivery and does not address availability attacks or metadata side channels.

Deployment compatibility. The mechanism requires a one-time token bundle to reach each user before reporting. In implementations that include a preprocessing step (Sec. 7.2), the tokens can be delivered through the same setup channel as the data-independent auxiliary inputs. Their one-time use and unlinkability ensure verification adds no user-identity linkage beyond what the shuffle model permits.

Overhead. Per accepted message, the analyzer performs one signature verification and one membership test in $\mathsf{Spent}_{\mathsf{eid}}$; per round, it stores at most $n(k+1)$ spent tokens.