

Bayesian Advantage of Re-Identification Attack in the Shuffle Model

Pengcheng Su
School of Computer Science
Peking University
pcs@pku.edu.cn

Haibo Cheng Ping Wang
National Engineering Research Center of Software Engineering
Peking University
hbcheng@pku.edu.cn, pwang@pku.edu.cn

Abstract—The shuffle model, which anonymizes data by randomly permuting user messages, has been widely adopted in both cryptography and differential privacy. In this work, we present the first systematic study of the Bayesian advantage in re-identifying a user’s message under the shuffle model. We begin with a basic setting: one sample is drawn from a distribution P , and $n - 1$ samples are drawn from a distribution Q , after which all n samples are randomly shuffled. We define $\beta_n(P, Q)$ as the success probability of a Bayes-optimal adversary in identifying the sample from P , and define the additive and multiplicative Bayesian advantages as $\text{Adv}_n^+(P, Q) = \beta_n(P, Q) - \frac{1}{n}$ and $\text{Adv}_n^\times(P, Q) = n \cdot \beta_n(P, Q)$, respectively.

We derive exact analytical expressions and asymptotic characterizations of $\beta_n(P, Q)$, along with evaluations in several representative scenarios. Furthermore, we establish (nearly) tight mutual bounds between the additive Bayesian advantage and the total variation distance.

Finally, we extend our analysis beyond the basic setting and present, for the first time, an upper bound on the success probability of Bayesian attacks in shuffle differential privacy. Specifically, when the outputs of n users—each processed through an ε -differentially private local randomizer—are shuffled, the probability that an attacker successfully re-identifies any target user’s message is at most e^ε/n .

Index Terms—shuffle model, re-identification attack, Bayesian advantage, differential privacy

I. INTRODUCTION

The shuffle model involves a trusted shuffler that receives messages from users, randomly permutes them, and then publishes the shuffled outputs in order to achieve anonymization [1], [2]. This model has demonstrated strong capabilities in both cryptography and differential privacy: it enables information-theoretically secure key exchange (which can be further applied to advanced tasks such as secure multiparty computation) [3], [4], and significantly amplifies the privacy guarantees of local differential privacy (LDP) mechanisms [2], [5]–[7].

In this paper, we initiate the first systematic study of a fundamental question in the shuffle model: suppose n users each independently generate a message y_i according to a probability distribution P_i for $i = 1, 2, \dots, n$. After applying the shuffler, the adversary observes the shuffled output $\mathbf{z} = \mathcal{S}(y_1, y_2, \dots, y_n) = (y_{\sigma(1)}, y_{\sigma(2)}, \dots, y_{\sigma(n)})$, where σ is

a random permutation¹. We ask: what is the probability that an adversary can correctly identify the message sent by a target user t —that is, recover the index $\sigma^{-1}(t)$?

The success probability of such an attack depends on the adversary’s prior knowledge. We assume a strong adversary who knows all distributions P_i . Under this assumption, the theoretically optimal strategy is to compute the posterior distribution via Bayes’ rule and select the most likely candidate [8].

We illustrate the notion of the re-identification attack in the shuffle model with a simple example. Consider an anonymous voting scenario: voter 1’s vote is known a priori to be either candidate A or candidate B, whereas every other voter is known *not* to vote for A or B. Even after the ballots are randomly permuted by the shuffler, an adversary can immediately locate voter 1’s ballot in the shuffled ballots with probability 1, because its observable value is unique. This example is trivial in the sense that the support of voter 1’s output distribution is disjoint from the supports of the other voters’ distributions.

More generally, if the other voters’ output distributions have nonempty overlap with voter 1’s distribution, the observations become ambiguous and the adversary faces nontrivial uncertainty. Analyzing this nontrivial regime is the main objective of this paper.

We begin by analyzing a basic setting in which the input distributions are specified as $P_1 = P$ and $P_2 = P_3 = \dots = P_n = Q$. This captures the scenario where a single sample drawn from P is hidden among $(n - 1)$ independently drawn samples from Q , and the adversary’s objective is to identify the one originating from P .

This setting is not only analytically tractable but also practically motivated: it forms the basis of so-called *honey techniques*, with the *Honeyword* system being a representative example (see Section 2 for details) [8]–[10]. In contrast, the general case—where the distributions $P_2, \dots, P_n$ are not necessarily identical—is typically analytically intractable [11]. Fortunately, we show that it can be reduced to this basic setting without loss of generality.

We define the success probability of the Bayes-optimal adversary in the basic setting as $\beta_n(P, Q)$. Since a trivial strategy

¹In this notation, $\sigma(i) = j$ means that the output originally from the j -th user is now placed in the i -th position. For example, if $\mathbf{z} = (y_2, y_4, \dots)$, then $\sigma(1) = 2$, $\sigma(2) = 4$, ...

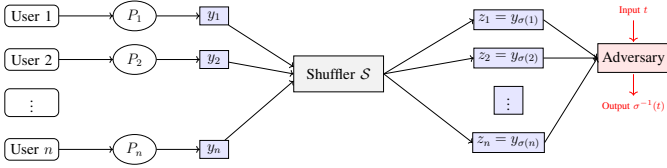


Fig. 1. Re-identification attack in the shuffle model. Each user i generates $y_i \sim P_i$, and the shuffler randomly permutes the outputs. The adversary observes $\mathbf{z} = (y_{\sigma(1)}, \dots, y_{\sigma(n)})$ and tries to infer the position of a target message.

that outputs a uniformly random index in $[n] = \{1, 2, \dots, n\}$ achieves a success probability of $1/n$, we define the *Bayesian advantage* in both additive and multiplicative forms as follows:

$$\text{Adv}_n^+(P, Q) = \beta_n(P, Q) - \frac{1}{n},$$

$$\text{Adv}_n^\times(P, Q) = \frac{\beta_n(P, Q)}{1/n} = n \cdot \beta_n(P, Q).$$

Leveraging the classical likelihood ratio method from information theory [12], we derive an exact analytical expression for $\beta_n(P, Q)$. This expression enables evaluation of $\beta_n(P, Q)$ in several representative cases and facilitates a precise characterization of its asymptotic behavior as n increases:

$$\beta_n(P, Q) = \alpha + \frac{M}{n} - \ell(n), \text{ where } \ell(n) \geq 0 \text{ and } \ell(n) = o\left(\frac{1}{n}\right),$$

with $M = \sup_{x: Q(x) \neq 0} \frac{P(x)}{Q(x)}$ and $\alpha = \sum_{x: Q(x) = 0} P(x)$. This result reveals that when P is absolutely continuous with respect to Q (i.e., $\alpha = 0$), the success probability is tightly upper bounded by $\frac{M}{n}$.

Furthermore, we establish a connection between $\text{Adv}_n^+(P, Q)$ and the classical total variation distance from information theory, defined as $\Delta(P, Q) = \frac{1}{2} \sum_x |P(x) - Q(x)|$. Specifically, we prove the following tight two-sided bound:

$$\frac{\Delta(P, Q)}{n} \leq \text{Adv}_n^+(P, Q) \leq \Delta(P, Q).$$

This result implies that for any fixed n , the additive Bayesian advantage $\text{Adv}_n^+(P, Q)$ is equivalent to $\Delta(P, Q)$ up to constant factors. In particular, as $\Delta(P, Q) \rightarrow 0$, the additive advantage $\text{Adv}_n^+(P, Q) \rightarrow 0$. We further demonstrate that both bounds are tight by providing explicit examples. The cryptographic implications of this connection are discussed in Section V.

Finally, we extend the analysis from the basic setting to the general setting. As a representative application, we consider the *single-message shuffle model*, a widely adopted paradigm in differential privacy [2], [5]–[7]. In this model, each user holds an input $x_i \in \mathbb{X}$ and applies a local randomizer $\mathcal{R}$ to produce a message $\mathcal{R}(x_i)$ ². These messages are then passed to a trusted shuffler, which applies a uniformly random permutation before forwarding the shuffled messages to the server.

² It is worth noting that our analysis allows the inputs x_i to be random variables on $\mathbb{X}$. A detailed discussion is provided in Section III-B.

Previous research has primarily focused on analyzing and computing the differential privacy guarantees in the shuffle model (see details in Section II-A) [2], [5]–[7], [13]. To the best of our knowledge, our work is the first to formally consider and theoretically analyze re-identification attacks in the Shuffle DP setting.

We prove that if the mechanism $\mathcal{R}$ satisfies ε -differential privacy—i.e., for all $x, x' \in \mathbb{X}$ and all outputs $y \in \mathbb{Y}$,

$$\frac{\Pr[\mathcal{R}(x) = y]}{\Pr[\mathcal{R}(x') = y]} \leq e^\varepsilon,$$

then the success probability of a Bayesian adversary in re-identifying a user's output is bounded by

$$\beta_n(\mathcal{R}) \leq \frac{e^\varepsilon}{n}.$$

This result indicates that in the shuffle DP model, the adversary's success probability in re-identifying any user's output is at most e^ε times the ideal baseline of $1/n$. In other words, the multiplicative Bayesian advantage $\text{Adv}_n^\times(\mathcal{R}) \leq e^\varepsilon$. A more refined analysis shows that when user 1 is targeted, the tight asymptotic bound on $\text{Adv}_n^\times(\mathcal{R}, x_1)$ is given by

$$\text{Adv}_n^\times(\mathcal{R}, x_1) \leq M,$$

where $M = \sup_y \frac{\Pr[\mathcal{R}(x_1) = y]}{\inf_x \Pr[\mathcal{R}(x) = y]}$.

Our main proof technique is based on a novel application of the two decomposition methods known as the *clone* [7] and the *blanket* [6]. These decompositions were originally developed to analyze privacy-amplification effects in the shuffle model; the reader may consult [14] for a survey of these methods. From a theoretical perspective, the reason decomposition techniques are well suited to analyzing both privacy amplification and re-identification attacks is that *both problems admit closed-form solutions in the basic setting*. In particular, Lemma 5.3 in the blanket paper [6] provides an analytic expression for the Hockey-Stick divergence of the shuffled outputs when user 1's input changes while all other users' messages are sampled i.i.d. from the same distribution; our basic-setting analysis for re-identification (Theorem 2) is analogous.

Intuitively, when the distributions of the other users are not identical, the decomposition approach discards the heterogeneous parts of those distributions and retains only the component that is common (i.e., the overlapping distribution). One can show that discarding the non-common parts can only increase the adversary's distinguishing power. Hence, applying these decompositions yields valid upper bounds on the shuffle mechanism's differential-privacy loss and on the Bayesian optimal re-identification success probability.

It is worth mentioning that we find the blanket decomposition to be optimal among all possible decompositions in the analysis of re-identification attacks. This parallels the result in [14], which established the optimality of the blanket decomposition for privacy amplification analysis. Overall, the clone decomposition, due to its simple form, provides consistent and concise bounds, whereas the blanket decomposition can yield tighter, mechanism-specific tailored bounds for particular differential privacy mechanisms.

Our contributions are summarized as follows:

- We introduce a formal definition of the Bayesian success probability $\beta_n(P, Q)$ for identifying a sample drawn from P among $n - 1$ decoys drawn from Q under random shuffling, and define the corresponding Bayesian advantage in both additive and multiplicative forms.
- We derive an exact analytical expression for $\beta_n(P, Q)$ using the likelihood ratio method, and characterize its asymptotic behavior.
- We establish mutual bounds connecting the Bayesian advantage and the total variation distance, showing that they are equivalent up to constant factors.
- We extend our analysis to the general shuffle model and prove that if the local randomizer $\mathcal{R}$ satisfies ε -DP, then the Bayesian re-identification success probability satisfies $\beta_n(\mathcal{R}) \leq \frac{e^\varepsilon}{n}$. We also derive a tight asymptotic bound on the multiplicative advantage.

II. BACKGROUND

A. Differential privacy

Differential privacy is a privacy-preserving framework for randomized algorithms. Intuitively, an algorithm is differentially private if the output distribution does not change significantly when a single individual's data is modified. This ensures that the output does not reveal substantial information about any individual in the dataset.

The classical definition of differential privacy (DP) assumes that the server has direct access to the data of all n users [15]. In real-world applications, *Local Differential Privacy* (LDP) is widely adopted, as it eliminates the need for a trusted curator by applying randomized noise to each user's data before any aggregation takes place [16]–[19]. However, this decentralized approach often incurs substantial utility loss due to the high level of noise required to ensure privacy.

To address this trade-off, the *shuffle model* introduces a trusted shuffler between the users and the aggregator [1], [2], [5]. In this work, we focus on the *single-message shuffle model*, in which each client sends a single report generated from their data, and these reports are anonymized by random shuffling before being forwarded to the server [13]. Here we give the formal definitions of DP, LDP, and shuffle DP.

We say that random variables P and Q are (ε, δ) -indistinguishable if for all set T :

$$\Pr[P \in T] \leq e^\varepsilon \Pr[Q \in T] + \delta.$$

If two datasets X and X' have the same size and differ only by the data of a single individual, they are referred to as neighboring datasets (denoted by $X \simeq X'$).

Definition 1 (Differential Privacy). *An algorithm $\mathcal{R} : \mathbb{X}^n \rightarrow \mathbb{Z}$ satisfies (ε, δ) -differential privacy if for all neighboring datasets $X, X' \in \mathbb{X}^n$, $\mathcal{R}(X)$ and $\mathcal{R}(X')$ are (ε, δ) -indistinguishable.*

Definition 2 (Local Differential Privacy). *An algorithm $\mathcal{R} : \mathbb{X} \rightarrow \mathbb{Y}$ satisfies local (ε, δ) -differential privacy if for all $x, x' \in \mathbb{X}$, $\mathcal{R}(x)$ and $\mathcal{R}(x')$ are (ε, δ) -indistinguishable.*

Here, ε is referred to as the privacy budget, which controls the privacy loss, while δ allows for a small probability of failure. When $\delta = 0$, the mechanism is also called ε -DP.

A single-message protocol $\mathcal{P}$ in the shuffle model is defined as a pair of algorithms $\mathcal{P} = (\mathcal{R}, \mathcal{A})$, where $\mathcal{R} : \mathbb{X} \rightarrow \mathbb{Y}$, and $\mathcal{A} : \mathbb{Y}^n \rightarrow \mathbb{O}$. We call $\mathcal{R}$ the *local randomizer*, $\mathbb{Y}$ the *message space* of the protocol, $\mathcal{A}$ the *analyzer*, and $\mathbb{O}$ the *output space* [2], [6].

The overall protocol implements a mechanism $\mathcal{P} : \mathbb{X}^n \rightarrow \mathbb{O}$ as follows: Each user i holds a data record x_i , to which they apply the local randomizer to obtain a message $y_i = \mathcal{R}(x_i)$. The messages y_i are then shuffled and submitted to the analyzer. Let $\mathcal{S}(y_1, \dots, y_n)$ denote the random shuffling step, where $\mathcal{S} : \mathbb{Y}^n \rightarrow \mathbb{Y}^n$ is a *shuffler* that applies a random permutation to its inputs.

In summary, the output of $\mathcal{P}(x_1, \dots, x_n)$ is given by

$$\mathcal{A} \circ \mathcal{S} \circ \mathcal{R}^n(\mathbf{x}) = \mathcal{A}(\mathcal{S}(\mathcal{R}(x_1), \dots, \mathcal{R}(x_n))).$$

Definition 3 (Differential Privacy in the Shuffle Model). *A protocol $\mathcal{P} = (\mathcal{R}, \mathcal{A})$ satisfies (ε, δ) -differential privacy in the shuffle model if for all neighboring datasets $X, X' \in \mathbb{X}^n$, the distributions $\mathcal{S} \circ \mathcal{R}^n(X)$ and $\mathcal{S} \circ \mathcal{R}^n(X')$ are (ε, δ) -indistinguishable.*

The “amplification-by-shuffling” theorem in the shuffle model implies that when each of the n users randomizes their data using an ε_0 -LDP mechanism, the collection of shuffled reports satisfies $(\varepsilon(\varepsilon_0, \delta, n), \delta)$ -DP, where $\varepsilon(\varepsilon_0, \delta, n) \ll \varepsilon_0$ for sufficiently large n and not too small δ [13].

Previous research has primarily focused on analyzing and computing the differential privacy guarantees—namely, the parameters (ε, δ) —achieved by a given protocol $\mathcal{P}$ in the shuffle model [2], [5]–[7], [13]. Our work is the first to formally consider and theoretically analyze re-identification attacks in the Shuffle DP setting.

B. Honeyword

Frequent password breaches—over 4,450 incidents exposing 28 billion records by July 2025 [20]—have motivated leakage detection mechanisms such as the Honeyword system [9].

In this system, the authentication server stores a user's real password alongside $n - 1$ plausible decoys—called *honeywords*—randomly shuffled to conceal the real one [9]. Upon a breach, the attacker obtains n indistinguishable candidates rather than the actual password. If a honeyword is later used in a login attempt, the system detects this anomaly and raises an alert [8], [9].

Let P denote the distribution of human-chosen passwords and Q the distribution of honeywords. The security of the Honeyword system is then captured by the Bayesian success probability $\beta_n(P, Q)$.

Prior studies on honeyword security typically consider an adversary who has access to a dataset D of human-chosen passwords, which is assumed to be sampled from the underlying distribution P , along with knowledge of the honeyword generation mechanism Q [8], [21]–[23]. Due to the lack of full

knowledge of P , the optimal attack strategy in this setting is analytically intractable. As a result, existing works rely on designing heuristic attack algorithms to evaluate the security of honeyword systems.

In contrast, we consider, for the first time, a theoretically strongest adversary who has complete knowledge of the distribution P . Under this setting, we provide an exact theoretical characterization of the Bayes-optimal success probability $\beta_n(P, Q)$, including its asymptotic analysis and its relationship to the total variation distance. We believe this perspective offers new insights into the security analysis of honeyword systems and related mechanisms.

III. PROBLEM FORMULATION

A. Re-identification attack in the basic setting

The formal definition of the re-identification attack in the basic setting is given in Game 1. Game 1 additionally introduces a parameter k , allowing the adversary $\mathcal{A}$ to submit k guesses. If the position of the sample drawn from distribution P (i.e., $\sigma^{-1}(1)$) is among the k guesses submitted by $\mathcal{A}$, then the adversary wins the game. We use the notation $y_1 \leftarrow_P \mathbb{Y}$ to denote that y_1 is sampled from distribution P over the domain $\mathbb{Y}$, and we write $\mathcal{P}_n$ to denote the set of all permutations over $[n]$.

Definition 4. We define the maximum success probability of such an adversary as

$$\beta_n^k(P, Q) = \max_{\mathcal{A}} \Pr \left[\text{GuessGame}_{\mathcal{A}}^{P, Q, n}(k) = \text{True} \right].$$

The special case $k = 1$, denoted $\beta_n^1(P, Q)$, corresponds to the single-guess setting, which is of primary interest in practice. For simplicity, we refer to it as $\beta_n(P, Q)$.

We consider an information-theoretic adversary: that is, an attacker with full knowledge of the distributions P and Q , and unbounded computational power. In Section IV-A, we show that the optimal attack strategy for such an adversary is to follow the posterior distribution induced by Bayes' rule. Consequently, we refer to $\beta_n(P, Q)$ as the *Bayesian success probability*.

Comparison with alternative definitions. It is important to emphasize that in our formulation, the adversary is tasked with identifying the position $\sigma^{-1}(1)$ of the sample drawn from P , rather than recovering its actual value y_1 . This distinction is essential, as defining success based on recovering the value y_1 leads to metrics that are not robust.

Consider, for instance, the trivial case where both P and Q deterministically output the same value (e.g., 1). In this scenario, an adversary that simply outputs $y_1 = 1$ will always succeed, achieving a success probability of 1.

As another example, let P and Q be uniform distributions over $[m]$. As m grows, the probability of value collisions in the multiset $\{y_1, \dots, y_n\}$ vanishes. Without such collisions, attempting to recover the position of the sample drawn from P reduces to uniform random guessing, yielding success probability exactly $1/n$.

Game 1: $\text{GuessGame}_{\mathcal{A}}^{P, Q, n}(k)$

```

1  $y_1 \leftarrow_P \mathbb{Y}$ 
2  $y_2, y_3, \dots, y_n \leftarrow_Q \mathbb{Y}$ 
3  $\mathbf{y} \leftarrow (y_1, y_2, y_3, \dots, y_n)$ 
4  $\sigma \leftarrow_{\$} \mathcal{P}_n$  // Sample a random permutation
5  $\mathbf{y}_{\sigma} \leftarrow (y_{\sigma(1)}, y_{\sigma(2)}, \dots, y_{\sigma(n)})$ 
6  $G = \{g_1, g_2, \dots, g_k\} \leftarrow \mathcal{A}(\mathbf{y}_{\sigma})$ 
7 return  $\sigma^{-1}(1) \in G$ .
```

Game 2: $\text{GuessGame}_{\mathcal{A}}^{\mathcal{R}, n}(x)$

```

1 for  $i = 1$  to  $n$  do
2    $\lfloor y_i \leftarrow_{\mathcal{R}(x_i)} \mathbb{Y}$ 
3  $\mathbf{y} \leftarrow (y_1, y_2, y_3, \dots, y_n)$ 
4  $\sigma \leftarrow_{\$} \mathcal{P}_n$  // Sample a random permutation
5  $\mathbf{y}_{\sigma} \leftarrow (y_{\sigma(1)}, y_{\sigma(2)}, \dots, y_{\sigma(n)})$ 
6  $g \leftarrow \mathcal{A}(\mathbf{y}_{\sigma})$ 
7 return  $g = \sigma^{-1}(1)$ .
```

By contrast, as formally shown in Section 4.1, when $P = Q$, the success probability of any adversary attempting to identify $\sigma^{-1}(1)$ is exactly $1/n$, regardless of the structure of P and Q . Furthermore, analyzing the success probability of recovering the actual value y_1 is often analytically intractable [11]. In comparison, $\beta_n(P, Q)$ admits a closed-form expression and is computationally tractable.

B. Re-identification attack in Shuffle DP

The formal definition of the re-identification attack in the shuffle DP setting is given in Game 2. For simplicity, we focus on the single-guess case; however, our analysis naturally extends to settings where the adversary is allowed multiple guesses.

For clarity, we first present the definitions of the Bayesian attack success probability and its associated advantage when $x_1, x_2, \dots, x_n$ take deterministic values. We then explain how these definitions can be naturally generalized to the case where each x_i is drawn from a probability distribution V_i .

Definition 5. The Bayesian re-identification success probability for a user with input x_1 in the Shuffle DP model is defined as

$$\beta_n(\mathcal{R}, x_1) = \max_{\mathcal{A}, \mathbf{x} \in \mathbb{X}^n, \mathbf{x}[1] = x_1} \Pr \left[\text{GuessGame}_{\mathcal{A}}^{\mathcal{R}, n}(\mathbf{x}) = \text{True} \right],$$

where the maximum is taken over all adversaries $\mathcal{A}$ and all input vectors $\mathbf{x}$ such that the first user's input is fixed to x_1 .

Definition 6. Define the worst-case Bayesian success probability in the Shuffle DP model as

$$\beta_n(\mathcal{R}) := \max_{x_1 \in \mathbb{X}} \beta_n(\mathcal{R}, x_1).$$

Definition 7. We define the corresponding Bayesian advantage in both additive and multiplicative forms:

$$\text{Adv}_n^+(\mathcal{R}, x_1) = \beta_n(\mathcal{R}, x_1) - \frac{1}{n},$$

$$\begin{aligned}\text{Adv}_n^\times(\mathcal{R}, x_1) &= \beta_n(\mathcal{R}, x_1) / \frac{1}{n} = n \cdot \beta_n(\mathcal{R}, x_1), \\ \text{Adv}_n^+(\mathcal{R}) &= \max_{x_1 \in \mathbb{X}} \text{Adv}_n^+(\mathcal{R}, x_1), \\ \text{Adv}_n^\times(\mathcal{R}) &= \max_{x_1 \in \mathbb{X}} \text{Adv}_n^\times(\mathcal{R}, x_1).\end{aligned}$$

In our theoretical analysis, we consider a powerful adversary who has full knowledge of the distributions $\mathcal{R}(x_i), i = 1, 2, \dots, n$. This strong assumption allows our results to upper-bound the capabilities of weaker adversaries who may only have partial information.

Interpretation. It is worth noting that the inputs x_i may themselves be random variables over $\mathbb{X}$. For instance, we may have $x_i \sim V_i$, where $V_i, i = 1, 2, \dots, n$, are probability distributions on $\mathbb{X}$. In this setting, we assume that the adversary knows each V_i and can additionally determine the corresponding output distribution $P_i = \mathcal{R}(x_i)$.

Let $\mathcal{D}[\mathbb{X}]$ denote the set of all probability distributions over $\mathbb{X}$. In the following, we distinguish between the notions of a random variable and a probability distribution: a probability distribution is a deterministic object, which can be described by a vector of probabilities; a random variable, on the other hand, is inherently stochastic and is said to follow a certain probability distribution.

It can be shown that for any probability distributions $V_1, V_2 \in \mathcal{D}[\mathbb{X}]$, its corresponding random variables v_1, v_2 and any ε -DP mechanism $\mathcal{R}$, the distributions $\mathcal{R}(v_1)$ and $\mathcal{R}(v_2)$ also satisfy ε -differential privacy:

$$\forall y: \frac{\Pr[\mathcal{R}(v_1) = y]}{\Pr[\mathcal{R}(v_2) = y]} \leq \frac{\sup_{x \in \mathbb{X}} \Pr[\mathcal{R}(x) = y]}{\inf_{x \in \mathbb{X}} \Pr[\mathcal{R}(x) = y]} \leq e^\varepsilon,$$

where we use $\Pr[\mathcal{R}(v_1) = y] = \sum_x \Pr[v_1 = x] \cdot \Pr[\mathcal{R}(x) = y] \in [\inf_{x \in \mathbb{X}} \Pr[\mathcal{R}(x) = y], \sup_{x \in \mathbb{X}} \Pr[\mathcal{R}(x) = y]]$, and the result follows directly from the definition of ε -DP.

Consequently, in the scenario where $x_i \sim V_i$, one can equivalently consider a new ε -DP mechanism

$$\mathcal{R}^*: \mathcal{D}[\mathbb{X}] \rightarrow \mathbb{Y}, \quad \mathcal{R}^*(V_i) = \mathcal{R}(x_i).$$

This perspective transforms the input from a random variable x_i into a deterministic probability distribution description V_i .

We illustrate the above transformation using the $\ln(3)$ -DP 2-ary randomized response mechanism [24].

Example III.1. Denote $\{1, 2, \dots, k\}$ by $[k]$ and the uniform distribution on $[k]$ by $\mathcal{U}_{[k]}$. For any $k \in \mathbb{N}$ and $\varepsilon > 0$, the k -randomized response mechanism $k\text{RR}: [k] \rightarrow [k]$ is defined as:

$$k\text{RR}(x) = \begin{cases} x, & \text{with probability } \frac{e^{\varepsilon_0} - 1}{e^{\varepsilon} + k - 1}, \\ y \sim \mathcal{U}_{[k]}, & \text{with probability } \frac{k}{e^{\varepsilon} + k - 1}. \end{cases}$$

If x_1 takes the fixed value 1, it corresponds to $V_1 = (1.0, 0)$. If x_1 takes value 1 with probability 0.4 and value 2 with probability 0.6, then $V_1 = (0.4, 0.6)$. For $V = (p, 1 - p)$, the mechanism $\mathcal{R}^*(V)$ defines a probability distribution over $\{1, 2\}$, with probabilities given by $(0.25 + 0.5p, 0.75 - 0.5p)$.

For notational convenience, we will henceforth write $\mathcal{R}(x_i)$; the reader may interpret this as $\mathcal{R}^*(V_i)$ in cases where $x_i \sim V_i$. For example, Definition 5 can be interpreted as follows:

Definition 8. The Bayesian re-identification success probability for a user with input $x_1 \sim V_i$ in the Shuffle DP model is defined as

$$\begin{aligned}\beta_n(\mathcal{R}, x_1) &:= \beta_n(\mathcal{R}^*, V_1) \\ &= \max_{\mathcal{A}, \mathbf{V} \in \mathcal{D}[\mathbb{X}]^n, \mathbf{V}[1] = V_1} \Pr \left[\text{GuessGame}_{\mathcal{A}}^{\mathcal{R}^*, n}(\mathbf{V}) = \text{True} \right],\end{aligned}$$

where the maximum is taken over all adversaries $\mathcal{A}$ and all input **distribution** vectors $\mathbf{V} = (V_1, V_2, \dots, V_n)$ such that the first user's input **distribution** is fixed to V_1 .

Since $\mathcal{R}^*$ still satisfies ε -DP, the clone decomposition remains valid. In addition, the blanket decomposition of $\mathcal{R}^*$ coincides with that of $\mathcal{R}$ (see Section VI-C for details). This justifies that our analysis applies without loss of generality.

C. Relation to quantitative information flow

This work focuses on re-identification attacks in the shuffle model. The most closely related line of research is *Quantitative Information Flow* (QIF) [25]. In the QIF framework, a system is modeled as an information-theoretic channel that takes in a secret input and produces an observable output. The amount of information leakage is defined as the difference between the *vulnerability* of the secret before and after passing through the channel—that is, the difference between the prior and posterior vulnerabilities, which quantify the adversary's ability to perform a successful inference attack [26], [27].

In other words, within the QIF formulation, the attacker aims to infer the value of the secret itself. In our terminology, this corresponds to guessing y_1 in Game 1 or x_1 in Game 2, whereas our formulation focuses on guessing the *position* of the secret (i.e., $\sigma^{-1}(1)$). We compare our definition with the QIF-based formulation of Game 1 in Section III-A.

There have been studies exploring QIF in the context of differential privacy [28], [29]; most notably, Reference [11] was the first to examine QIF in the shuffle DP setting. Interestingly, they compare the QIF of LDP, shuffle-only, and combined LDP+shuffle settings, and observe that, under an uninformed adversary (who does not know any individual's data prior to accessing a dataset and assumes a uniform prior over datasets), the shuffling process contributes the majority of the privacy in k -RR. However, QIF analysis in general is known to be technically challenging, and closed-form expressions are available only for a few special cases [11].

Specifically, their work focuses on an *uninformed adversary*. In contrast, prior research typically considers a *strong adversary*, who knows all individuals' data except that of the target user. They provide an analytical treatment of the quantitative information flow for k -RR against an uninformed adversary, whereas the QIF of other mechanisms under an uninformed adversary remains unclear. In the case of a strong adversary, even the simple k -RR mechanism is analytically intractable; they only provide a brief discussion based on

numerical experiments and leave a thorough analysis for future work.

In contrast, our work demonstrates that for re-identification attacks, the Bayesian advantage in shuffle DP admits a much simpler and unified analysis.

IV. ANALYSIS IN THE BASIC SETTING

A. Bayesian optimal attack

We now derive the success probability of an adversary $\mathcal{A}$ in the basic setting, where the distributions P and Q are known, and the adversary observes a randomly shuffled vector $\mathbf{y}_\sigma = \mathbf{y}'$ and outputs a guess g for the index of the sample drawn from P .

$$\begin{aligned}
\Pr[\sigma^{-1}(1) = g \mid \mathbf{y}_\sigma = \mathbf{y}'] &= \Pr[\sigma^{-1}(1) = g \mid \mathbf{y} = \mathbf{y}'_{\sigma^{-1}}] \\
&= \frac{\Pr[\sigma^{-1}(1) = g \wedge \mathbf{y} = \mathbf{y}'_{\sigma^{-1}}]}{\Pr[\mathbf{y} = \mathbf{y}'_{\sigma^{-1}}]} \\
&= \frac{\Pr[\sigma^{-1}(1) = g] \cdot \Pr[\mathbf{y} = \mathbf{y}'_{\sigma^{-1}} \mid \sigma^{-1}(1) = g]}{\sum_{\sigma_1 \in \mathcal{P}_n} \Pr[\sigma = \sigma_1] \cdot \Pr[\mathbf{y} = \mathbf{y}'_{\sigma^{-1}} \mid \sigma = \sigma_1]} \\
&= \frac{\frac{1}{n} \cdot \frac{P(y'_g)}{Q(y'_g)} \cdot \prod_{j=1}^n Q(y'_j)}{\sum_{i=1}^n \frac{1}{n} \cdot \frac{P(y'_i)}{Q(y'_i)} \cdot \prod_{j=1}^n Q(y'_j)} \\
&= \frac{\frac{P(y'_g)}{Q(y'_g)}}{\sum_{i=1}^n \frac{P(y'_i)}{Q(y'_i)}}.
\end{aligned} \tag{1} \tag{2} \tag{3} \tag{4} \tag{5}$$

Here, σ is a uniformly random permutation over $[n] = \{1, 2, \dots, n\}$. Equation (1) holds because for each i , $y_i = y'_{\sigma^{-1}(i)}$. Equation (2) follows from the definition of conditional probability, and Equation (3) expands the denominator using the law of total probability.

Equation (4) leverages two facts: (i) $\Pr[\sigma^{-1}(1) = g] = 1/n$ due to uniform randomness, and (ii) conditioned on $\sigma^{-1}(1) = g$, the value y'_g is drawn from P , while all other y'_j are drawn from Q . This gives:

$$\begin{aligned}
\Pr[\mathbf{y} = \mathbf{y}'_{\sigma^{-1}} \mid \sigma^{-1}(1) = g] &= P(y'_g) \cdot \prod_{j \neq g} Q(y'_j) \\
&= \frac{P(y'_g)}{Q(y'_g)} \cdot \prod_{j=1}^n Q(y'_j).
\end{aligned}$$

In detail, denote $\mathcal{P}'_n = \{\sigma \mid \sigma^{-1}(1) = g\}$. Conditioning on $\sigma^{-1}(1) = g$ implies that σ is uniformly sampled from $\mathcal{P}'_n$, where $|\mathcal{P}'_n| = (n-1)!$. For any $\sigma_* \in \mathcal{P}'_n$, we have

$$\Pr[\mathbf{y} = \mathbf{y}'_{\sigma^{-1}} \mid \sigma = \sigma_*] = \Pr[\mathbf{y} = \mathbf{y}'_{\sigma_*^{-1}}] = P(y'_g) \cdot \prod_{j \neq g} Q(y'_j),$$

which is independent of the choice of σ_* . Consequently,

$$\begin{aligned}
\Pr[\mathbf{y} = \mathbf{y}'_{\sigma^{-1}} \mid \sigma^{-1}(1) = g] &= \sum_{\sigma_* \in \mathcal{P}'_n} \frac{1}{(n-1)!} \Pr[\mathbf{y} = \mathbf{y}'_{\sigma_*^{-1}}] \\
&= P(y'_g) \cdot \prod_{j \neq g} Q(y'_j).
\end{aligned}$$

The same logic applies to each term in the denominator.

Equation (5) reveals the form of the **Bayes-optimal strategy**: the adversary computes the likelihood ratio $P(y'_i)/Q(y'_i)$ for all $i \in [n]$, and ranks them in descending order. The index g with the largest ratio is the most likely to be the sample from P .

Importantly, if for some i , $Q(y'_i) = 0$ but $P(y'_i) > 0$, we interpret $P(y'_i)/Q(y'_i) = \infty$, indicating that y'_i must have come from P with certainty. In that case, the attacker can identify the position $\sigma^{-1}(1) = i$ with probability 1.

Corollary 1. *If $P = Q$, then for any $n \geq 1$, the success probability of any adversary $\mathcal{A}$ in the basic setting is exactly $1/n$. That is,*

$$\beta_n(P, P) = \frac{1}{n}, \quad \text{Adv}_n^+(P, P) = 0.$$

Proof. From Equation (5), when $P = Q$, the likelihood ratio $P(y'_i)/Q(y'_i) = 1$ for all $i \in [n]$. Therefore, the adversary sees n identical scores, and no position is statistically distinguishable from the others. Hence, any strategy reduces to uniform guessing, and the success probability is $1/n$. $\square$

B. Theoretical derivation of $\beta_n(P, Q)$

As we will demonstrate, the core of the theoretical computation of $\beta_n(P, Q)$ lies in an insightful shift in perspective. The crucial idea is to transition from working directly with the two distributions P and Q to studying the distributions of their likelihood ratio $\frac{P(y)}{Q(y)}$.

We define two cumulative distribution functions:

$$F(t) = \Pr_{y \sim P} \left[\frac{P(y)}{Q(y)} \leq t \right], \tag{6a}$$

$$G(t) = \Pr_{y \sim Q} \left[\frac{P(y)}{Q(y)} \leq t \right], \tag{6b}$$

where $\frac{P(y)}{Q(y)}$ is the likelihood ratio of y , a quantity with broad applications in information theory [12]. Let $f(t)$ and $g(t)$ denote the corresponding probability density functions of F and G , respectively.

We also introduce two useful notations³:

$$f(+\infty) := \Pr_{y \sim P}[Q(y) = 0], \quad M := \sup_{y: Q(y) \neq 0} \frac{P(y)}{Q(y)}.$$

Recall that the Bayes-optimal attacker sorts the entries by descending likelihood ratio. The distribution of the likelihood ratio of a sample drawn from P is exactly f . For example, if the likelihood ratio of y_1 is 1.4, then the attacker will succeed in the first guess if and only if all other y_i , $i = 2, 3, \dots, n$, have likelihood ratios less than 1.4. The probability of this event is $f(1.4) \cdot G^{n-1}(1.4)$. Integrating over all possible values of the likelihood ratio yields:

$$\beta_n(P, Q) = f(+\infty) + \int_0^M f(t) G^{n-1}(t) dt. \tag{7}$$

A similar approach yields the expression for $\beta_n^k(P, Q)$. By the definition of $\beta_n^k(P, Q)$ and the optimal strategy, the

³According to the definition, M may be $+\infty$ in some cases.

Bayesian adversary wins in $\text{GuessGame}_A^{P,Q,n}(k)$ if and only if the number of y_i ($i = 2, \dots, n$) with likelihood ratios greater than that of y_1 is fewer than k .

Theorem 1. *The Bayesian success probability with k guesses is given by:*

$$\beta_n^k(P, Q) = f(+\infty) + \sum_{j=1}^k \int_0^M \binom{n-1}{j-1} f(t) G^{n-j}(t) (1-G(t))^{j-1} dt. \quad (8)$$

Proof. The adversary succeeds in the j -th guess if and only if exactly $j-1$ of the $n-1$ decoy samples have likelihood ratios greater than that of y_1 . The integrand accounts for all configurations where this occurs, and the integral aggregates over all possible threshold values t . Summing over j from 1 to k gives the desired result. $\square$

The following property will be essential in simplifying these expressions:

Property 1. *The density functions f and g defined above satisfy:*

$$f(t) = t \cdot g(t). \quad (9)$$

Proof. Let $S_t = \{y \in \mathbb{Y} \mid \frac{P(y)}{Q(y)} = t\}$. Then:

$$\frac{f(t)}{g(t)} = \frac{P(S_t)}{Q(S_t)} = \frac{\sum_{y \in S_t} P(y)}{\sum_{y \in S_t} Q(y)} = \frac{\sum_{y \in S_t} t \cdot Q(y)}{\sum_{y \in S_t} Q(y)} = t. \quad \square$$

Theorem 2. *Let P and Q be probability distributions such that*

$$M := \sup_{y: Q(y) \neq 0} \frac{P(y)}{Q(y)} < \infty.$$

Then the Bayesian success probability $\beta_n(P, Q)$ can be expressed as:

$$\beta_n(P, Q) = \frac{1}{n} \left(M - \int_0^M G^n(t) dt \right) + f(+\infty). \quad (10)$$

Proof. Using integration by parts and Property 1:

$$\begin{aligned} \beta_n(P, Q) &= \int_0^M f(t) G^{n-1}(t) dt + f(+\infty) \\ &= \int_0^M t \cdot g(t) \cdot G^{n-1}(t) dt + f(+\infty) \\ &= \int_0^M t \cdot G^{n-1}(t) dG(t) + f(+\infty) \\ &= \frac{1}{n} \int_0^M t dG^n(t) + f(+\infty) \\ &= \frac{1}{n} \left[t G^n(t) \Big|_0^M - \int_0^M G^n(t) dt \right] + f(+\infty) \\ &= \frac{1}{n} \left(M - \int_0^M G^n(t) dt \right) + f(+\infty). \quad \square \end{aligned}$$

Equation (10) clearly shows how $\beta_n(P, Q)$ scales with n :

Theorem 3. *Let P and Q be probability distributions such that $M := \sup_{y: Q(y) \neq 0} \frac{P(y)}{Q(y)} < \infty$. Then the Bayesian success probability*

$$\beta_n(P, Q) = f(+\infty) + \frac{M}{n} - \ell(n),$$

where $\ell(n) \geq 0$ and $\ell(n) = o(\frac{1}{n})$.

Proof. From Theorem 2, we have

$$\beta_n(P, Q) = f(+\infty) + \frac{M}{n} - \frac{1}{n} \int_0^M G^n(t) dt.$$

Define $\ell(n) := \frac{1}{n} \int_0^M G^n(t) dt$. Since $0 \leq G(t) \leq 1$ for all t (as G is a cumulative distribution function), it follows that $0 \leq \ell(n) \leq \frac{M}{n}$. Moreover, as $G^n(t) \rightarrow 0$ pointwise for $t < M$ when $n \rightarrow \infty$, and the integrand is dominated by an integrable function, the dominated convergence theorem implies:

$$\ell(n) = \frac{1}{n} \int_0^M G^n(t) dt = o\left(\frac{1}{n}\right). \quad \square$$

There is a subtle point in the derivation of $\beta_n(P, Q)$ that warrants clarification: when multiple entries y_i share the same maximum likelihood ratio, how should ties be resolved? This directly determines whether the cumulative function $G(t)$ in the expression $f(t)G^{n-1}(t)$ includes the probability mass at t .

We formalize this subtlety in the following result:

Theorem 4. *Let P and Q be discrete probability distributions. Then the Bayesian success probability is given by:*

$$\beta_n(P, Q) = f(+\infty) + \frac{1}{n} \sum_t t [G^n(t) - G^n(t^-)],$$

where $G(t^-)$ denotes the left-hand limit of G at t , i.e., $G(t^-) = \Pr_{y \sim Q} \left[\frac{P(y)}{Q(y)} < t \right]$.

The proof is provided in Appendix A, which proceeds via a combinatorial calculation. This result is the discrete analogue of the continuous case:

$$\beta_n(P, Q) = f(+\infty) + \frac{1}{n} \int_0^M t dG^n(t).$$

C. Case studies

In this section, we demonstrated how to compute $\beta_n(P, Q)$ under two representative examples. In both cases, the decoy distribution Q is chosen to be uniform—this choice is made not only to facilitate theoretical analysis, but also because the uniform distribution is commonly adopted in practice when little is known about the target distribution P .

Example IV.1. *Consider the following two continuous probability distributions defined on the interval $[0, 1]$:*

$$\begin{aligned} P(x) &= x + 0.5, & 0 \leq x \leq 1, \\ Q(x) &= 1, & 0 \leq x \leq 1. \end{aligned}$$

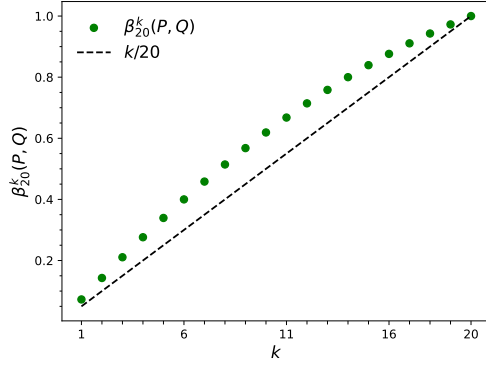


Fig. 2. $\beta_n^k(P, Q)$ as a function of k for $n = 20$ in Example IV.1

In this case, the likelihood ratio $\frac{P(x)}{Q(x)} = x + 0.5$, and it follows:

$$\begin{aligned} f(t) &= t, & 0.5 \leq t \leq 1.5, \\ g(t) &= 1, & 0.5 \leq t \leq 1.5. \end{aligned}$$

Using the formula in Theorem 2, we compute:

$$\begin{aligned} \beta_n(P, Q) &= \frac{1}{n} \left(M - \int_0^M G^n(t) dt \right) = \frac{1}{n} \left(1.5 - \frac{1}{n+1} \right) \\ &= \frac{1.5}{n} - \frac{1}{n(n+1)}. \end{aligned}$$

Hence, $\beta_n(P, Q) = \Theta(1.5/n)$.

We can also compute the generalized success probability for k guesses:

$$\beta_n^k(P, Q) = \frac{3n+2}{2n(n+1)} \cdot k - \frac{k^2}{2n(n+1)}.$$

This is a quadratic function of k . The function $\beta_n^k(P, Q)$ versus k for $n = 20$ is illustrated in Figure 2.

The next example involves the Honeyword setting. To model real-world password behavior, we assume a non-uniform distribution P for real passwords. Empirical studies have shown that human-chosen passwords approximately follow a Zipf distribution [30], [31], where the r -th most probable password has probability:

$$P(pw_r) = \frac{r^{-\alpha}}{\sum_{i=1}^m i^{-\alpha}},$$

with parameter $0 < \alpha < 1$ and $m = |\mathcal{PW}|$ denoting the size of the password space. This reflects the intuition that a small number of high-ranked passwords dominate the distribution.

Example IV.2. Let P be a Zipf distribution with parameter α over $[m]$, and let Q be the uniform distribution over $[m]$. Define $S = \sum_{j=1}^m j^{-\alpha} \approx \frac{1}{1-\alpha} m^{1-\alpha}$. Then, the Bayesian success probability $\beta_n(P, Q)$ is given by:

$$\beta_n(P, Q) = \frac{1}{n} \sum_t t [G^n(t) - G^n(t^-)]$$

$$= \frac{1}{n} \sum_{i=1}^m m \cdot \frac{(m-i+1)^{-\alpha}}{S} \left[\left(\frac{i}{m} \right)^n - \left(\frac{i-1}{m} \right)^n \right],$$

where we use the fact that under P and Q , the likelihood ratio $\frac{P(i)}{Q(i)}$ equals $m \cdot \frac{(m-i+1)^{-\alpha}}{S}$.

To simplify this expression, we approximate the finite difference by a derivative:

$$\left(\frac{i}{m} \right)^n - \left(\frac{i-1}{m} \right)^n \approx \frac{n}{m} \left(\frac{i}{m} \right)^{n-1},$$

and interpret the summation as a Riemann sum:

$$\begin{aligned} \beta_n(P, Q) &\approx \frac{m^{1-\alpha}}{S} \sum_{i=1}^m \frac{1}{m} \left(1 - \frac{i+1}{m} \right)^{-\alpha} \left(\frac{i}{m} \right)^{n-1} \\ &\approx \frac{m^{1-\alpha}}{S} \int_0^1 (1-t)^{-\alpha} t^{n-1} dt. \end{aligned}$$

This integral corresponds to the Beta function:

$$\beta_n(P, Q) \approx \frac{m^{1-\alpha}}{S} \cdot B(1-\alpha, n),$$

where

$$B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt$$

is the Beta function. Substituting the approximation of S , we obtain:

$$\beta_n(P, Q) \approx (1-\alpha) \cdot B(1-\alpha, n).$$

Likewise, we can compute the multi-guess success probability:

$$\beta_n^k(P, Q) = \sum_{j=1}^k (1-\alpha) \cdot \binom{n-1}{j-1} \cdot B(j-\alpha, n+1-j).$$

Assuming $\alpha = 0.7$ [21], the plot of $\beta_n(P, Q)$ versus n is shown in Figure 3. The graph of $\beta_n^k(P, Q)$ as a function of k for $n = 20$ is shown in Figure 4. Numerical evaluations show that when P is a Zipf distribution with parameter 0.7 and Q is uniform, the Bayes-optimal adversary achieves a success probability $\beta_n(P, Q) < 0.2$ only if $n \geq 150$. Moreover, for $n = 20$, the success probability of the optimal adversary within the first $k = 3$ guesses, i.e., $\beta_n^3(P, Q)$, already exceeds 0.5.

Our theoretical results offer quantitative guidance for estimating the attack success probability of a Bayes-optimal adversary in this setting and for selecting the number of decoys necessary to ensure adequate security.

V. CONNECTIONS BETWEEN ADDITIVE BAYESIAN ADVANTAGE AND TOTAL VARIATION

A. Main results

Many cryptographic definitions implicitly rely on distinguishing games between two probability distributions. A classical setting is as follows: the challenger samples $m_0 \sim P$ and $m_1 \sim Q$, selects a random bit $b \in \{0, 1\}$, and sends m_b to the adversary, who must then guess the value of b by outputting b' . The adversary's success probability is $\Pr[b' = b]$, and it is

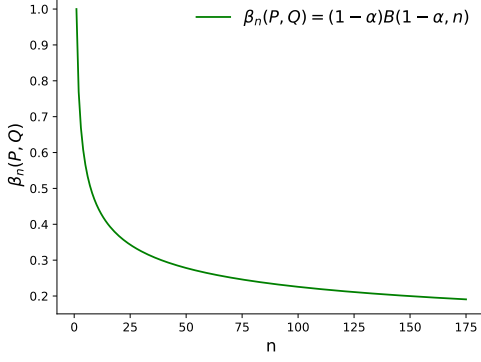


Fig. 3. $\beta_n(P, Q)$ vs. n where $P = \text{Zipf}(0.7)$ and Q is uniform

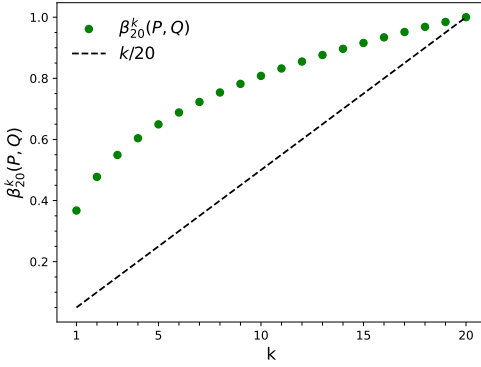


Fig. 4. $\beta_{20}^k(P, Q)$ vs. k where $P = \text{Zipf}(0.7)$ and Q is uniform

well known that the maximal advantage over random guessing is characterized by:

$$\max_{\mathcal{A}} \Pr[b' = b] = \frac{1}{2} + \frac{1}{2} \Delta(P, Q),$$

where $\Delta(P, Q) := \frac{1}{2} \sum_{x \in \mathbb{X}} |P(x) - Q(x)|$ denotes the total variation distance between P and Q .

Now consider a stronger challenge where both m_0 and m_1 are given to the adversary, who must decide which one came from P and which from Q . This corresponds to our re-identification setting $\text{GuessGame}_{\mathcal{A}}^{P, Q, n}(k)$ with $n = 2$ and $k = 1$. In this case, the optimal success probability is given by $\beta_2(P, Q)$, and the Bayesian advantage becomes $\text{Adv}_2^+(P, Q) = \beta_2(P, Q) - \frac{1}{2}$.

The following theorem establishes a tight relationship between $\text{Adv}_n^+(P, Q)$ and the total variation distance $\Delta(P, Q)$.

Theorem 5. *For any distributions P, Q and integer $n \geq 1$, the Bayesian advantage satisfies:*

$$\frac{1}{n} \cdot \Delta(P, Q) \leq \text{Adv}_n^+(P, Q) \leq \Delta(P, Q).$$

The proof is deferred to Appendix B. The lower bound is established via an explicit construction of an adversary's

strategy, whereas the upper bound is more involved and requires several analytical techniques.

B. Tightness analysis

Tightness of the upper bound. Consider the distributions $P(0) = p$, $P(1) = 1 - p$ and $Q(0) = 0$, $Q(1) = 1$. In this case, the total variation distance is $\Delta(P, Q) = p$, and a direct computation using Theorem 4 yields:

$$\beta_n(P, Q) = p + \frac{1-p}{n}, \quad \text{Adv}_n^+(P, Q) = \left(1 - \frac{1}{n}\right)p.$$

This demonstrates that for any fixed total variation distance $\Delta(P, Q) = p$, there exist distributions P and Q such that:

$$\text{Adv}_n^+(P, Q) = \left(1 - \frac{1}{n}\right) \Delta(P, Q).$$

Therefore, the upper bound $\text{Adv}_n^+(P, Q) \leq \Delta(P, Q)$ is tight up to a factor of $1/n$, and cannot be improved in general.

Tightness of the lower bound. Consider the distributions $P(0) = 0$, $P(1) = 1$, and $Q(0) = p$, $Q(1) = 1 - p$. Then the total variation distance is $\Delta(P, Q) = p$. A direct computation shows:

$$\text{Adv}_n^+(P, Q) = \frac{1}{(1-p)n} (1-p^n) - \frac{1}{n} = \frac{p}{(1-p)n} - \frac{p^n}{(1-p)n}.$$

For any fixed $n \geq 2$, as $p \rightarrow 0$, we have $\text{Adv}_n^+(P, Q) = \Theta\left(\frac{p}{n}\right)$, which matches the lower bound $\frac{1}{n} \Delta(P, Q)$ up to a vanishing multiplicative factor.

VI. ANALYSIS OF BAYESIAN ADVANTAGE IN SHUFFLE DP

A. Reduction to the basic setting

As discussed earlier, unlike the basic setting—where the Bayesian attack admits a closed-form analytical expression—the primary challenge in analyzing the shuffle DP model stems from the fact that the output distributions $\mathcal{R}(x_i)$ for users $i = 2, 3, \dots, n$ are generally non-identical.

Fortunately, this difficulty can be circumvented by reducing the shuffle DP analysis to the basic setting through a probabilistic decomposition of the local randomizer $\mathcal{R}$. Specifically, suppose $\mathcal{R}$ satisfies the following mixture structure:

$$\forall x \in \mathbb{X}: \quad \mathcal{R}(x) = \gamma Q^{\text{com}} + (1 - \gamma) \cdot \text{LO}(x),$$

where Q^{com} is a distribution shared across all users (the *common component*), and $\text{LO}(x)$ is an input-dependent *left-over* distribution.

Under this decomposition, the re-identification problem reduces to the basic setting with $P = \mathcal{R}(x_1)$, $Q = Q^{\text{com}}$, and a random number of decoys $N \sim 1 + \text{Bin}(n-1, \gamma)$, where $\text{Bin}(n-1, \gamma)$ denotes the binomial distribution with $n-1$ trials and success probability γ . This reformulation is formalized in Game 3.

We define:

$$\psi(\mathcal{R}, n, \gamma, Q^{\text{com}}, x_1) :=$$

$$\max_{\mathcal{A}} \Pr \left[\text{GuessGame}_{\mathcal{A}}^{\mathcal{R}, n, \gamma, Q^{\text{com}}}(x_1) = \text{True} \right].$$

Game 3: $\text{GuessGame}_{\mathcal{A}}^{\mathcal{R},n,\gamma,Q^{\text{com}}}(x_1)$

```

1  $y_1 \leftarrow \mathcal{R}(x_1) \mathbb{Y}$ 
2  $N \sim 1 + \text{Bin}(n-1, \gamma)$ 
3  $y_2, y_3, \dots, y_N \leftarrow Q^{\text{com}} \mathbb{Y}$ 
4  $\mathbf{y} \leftarrow (y_1, y_2, \dots, y_N)$ 
5  $\sigma \leftarrow_{\$} \mathcal{P}_N$  // Sample a random permutation
6  $\mathbf{y}_{\sigma} \leftarrow (y_{\sigma(1)}, y_{\sigma(2)}, \dots, y_{\sigma(N)})$ 
7  $g \leftarrow \mathcal{A}(\mathbf{y}_{\sigma})$ 
8 return  $\sigma^{-1}(1) = g$ 

```

Theorem 6. If the local randomizer $\mathcal{R}$ admits a decomposition of the form

$$\forall x \in \mathbb{X}: \quad \mathcal{R}(x) = \gamma Q^{\text{com}} + (1 - \gamma) \cdot \text{LO}(x),$$

then

$$\beta_n(\mathcal{R}, x_1) \leq \psi(\mathcal{R}, n, \gamma, Q^{\text{com}}, x_1).$$

Proof. Let $\mathbf{x}_{-1} = (x_2, x_3, \dots, x_n)$ denote the inputs of users 2 to n , and let $\mathbf{y} = (y_1, \dots, y_n)$ with $y_i \sim \mathcal{R}(x_i)$ be their corresponding outputs. Define a binary vector $\mathbf{b} = (b_2, \dots, b_n) \in \{0, 1\}^{n-1}$ such that $b_i = 1$ indicates $y_i \sim Q^{\text{com}}$, and $b_i = 0$ indicates $y_i \sim \text{LO}(x_i)$. Let $I = \{i \in [2, n] \mid b_i = 0\}$ be the index set of outputs not sampled from Q^{com} .

Game 4 corresponds to a variant of Game 2 in which the adversary is additionally provided the positions $\sigma^{-1}(I)$ of samples drawn from the non-common distributions $\text{LO}(x_i)$. We proceed by a hybrid argument. If the adversary is given $\sigma^{-1}(I)$, then the optimal strategy reduces to that in Game 3. If not, the adversary operates as in Game 2.

Let $\mathcal{A}$ be any adversary in the original game $\text{GuessGame}_{\mathcal{A}}^{\mathcal{R},n}(x)$. We define a new adversary $\mathcal{B}$ for Game 4 that ignores the extra information $\sigma^{-1}(I)$ and simply invokes $\mathcal{A}$ on the permuted outputs $\mathbf{y}_{\sigma}$. Since $\mathcal{A}$ does not utilize the additional information, the success probability remains unchanged. Hence,

$$\begin{aligned} & \forall \mathcal{A}, \exists \mathcal{B}, \forall \mathbf{x} \in \{\mathbf{x} \in \mathbb{X}^n \mid \mathbf{x}[1] = x_1\}, \\ & \Pr[\text{GuessGame}_{\mathcal{A}}^{\mathcal{R},n}(\mathbf{x}) = \text{True}] \\ &= \Pr[\text{GuessGame2}_{\mathcal{B}}^{\mathcal{R},n,\gamma,Q^{\text{com}}}(\mathbf{x}_1) = \text{True}]. \end{aligned}$$

Taking the maximum over all adversaries $\mathcal{A}$ and corresponding $\mathcal{B}$, we obtain:

$$\beta_n(\mathcal{R}, x_1) \leq \psi(\mathcal{R}, n, \gamma, Q^{\text{com}}, x_1). \quad \square$$

Theorem 7. The Bayesian success probability in the reduced game satisfies:

$$\begin{aligned} & \psi(\mathcal{R}, n, \gamma, Q^{\text{com}}, x_1) \\ &= \sum_{m=0}^{n-1} \binom{n-1}{m} \gamma^m (1-\gamma)^{n-1-m} \cdot \beta_{m+1}(\mathcal{R}(x_1), Q^{\text{com}}). \end{aligned}$$

Proof. This follows directly from the definition of $\psi(\cdot)$ in Game 3 and the definition of $\beta_n(P, Q)$.

Game 4: $\text{GuessGame2}_{\mathcal{A}}^{\mathcal{R},n,\gamma,Q^{\text{com}}}(x_1)$

```

1  $y_1 \leftarrow \mathcal{R}(x_1) \mathbb{Y}$ 
2  $I \leftarrow \emptyset$ 
3 for  $i = 2$  to  $n$  do
4    $b_i \sim \text{Bern}(\gamma)$ 
5   if  $b_i = 1$  then
6      $y_i \leftarrow Q^{\text{com}} \mathbb{Y}$ 
7   else
8      $y_i \leftarrow \text{LO}(x_i) \mathbb{Y}; \quad I \leftarrow I \cup \{i\}$ 
9  $\mathbf{y} \leftarrow (y_1, y_2, \dots, y_n)$ 
10  $\sigma \leftarrow_{\$} \mathcal{P}_n$  // Sample a random permutation
11  $\mathbf{y}_{\sigma} \leftarrow (y_{\sigma(1)}, y_{\sigma(2)}, \dots, y_{\sigma(n)})$ 
12  $g \leftarrow \mathcal{A}(\mathbf{y}_{\sigma}, \sigma^{-1}(I))$ 
13 return  $\sigma^{-1}(1) = g$ 

```

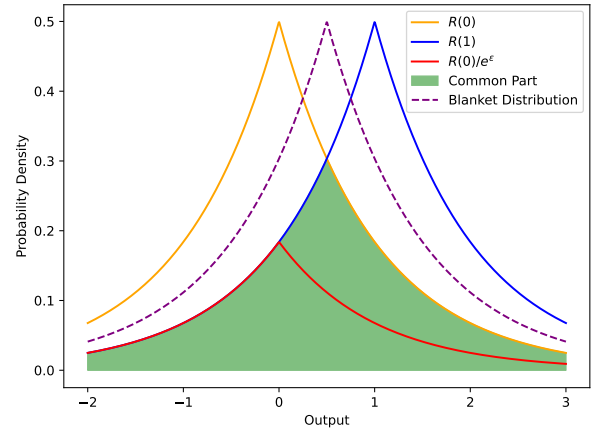


Fig. 5. Decomposition methods for 1-DP Laplace mechanism [15] on $\{0, 1\}$

In Game 3, the total number of samples drawn from Q^{com} (excluding the one from $P = \mathcal{R}(x_1)$) is a random variable $m \sim \text{Bin}(n-1, \gamma)$. For each such m , the attacker's success probability is $\beta_{m+1}(\mathcal{R}(x_1), Q^{\text{com}})$, since there are m decoys and one target. Taking the expectation over the binomial distribution gives the result. $\square$

Possible decompositions. The decomposition allowed by Theorem 6 admits multiple choices, among which the most commonly used are the *clone* and *blanket* decompositions, which we introduce separately below. For every ε -DP local randomizer, both of these decompositions are guaranteed to exist.

For an (ε, δ) -DP local randomizer $\mathcal{R}$, the clone decomposition may fail to exist (i.e., $\gamma = 0$ when $Q^{\text{com}} = \mathcal{R}(x_1)$). The blanket decomposition, on the other hand, always exists but generally lacks a well-structured form, which introduces additional challenges. We leave the analysis of this scenario as future work.

B. Analysis via the clone decomposition

The *clone* technique, originally proposed in [7] for analyzing privacy amplification in shuffle DP, can also be applied to the analysis of re-identification attacks in the shuffle model. Specifically, for any ε -DP local randomizer $\mathcal{R}$, the clone method ensures the following decomposition:

$$\forall x \in \mathbb{X}, \quad \mathcal{R}(x) = e^{-\varepsilon} \cdot \mathcal{R}(x_1) + (1 - e^{-\varepsilon}) \cdot \text{LO}(x).$$

This decomposition follows directly from the ε -DP property, which implies that for all $x \in \mathbb{X}$ and $y \in \mathbb{Y}$,

$$\Pr[\mathcal{R}(x) = y] \geq e^{-\varepsilon} \cdot \Pr[\mathcal{R}(x_1) = y],$$

thereby ensuring the existence of $\text{LO}(x_i)$ as a valid residual distribution.

Theorem 8. *Let $\mathcal{R}$ be an ε -differentially private local randomizer. Then, in the single-message shuffle model, the Bayesian success probability of any re-identification attack (as defined in Game 2) satisfies:*

$$\beta_n(\mathcal{R}) \leq \frac{e^\varepsilon}{n}.$$

Proof. By Theorem 6, using the *clone decomposition*, we have:

$$\beta_n(\mathcal{R}, x_1) \leq \psi(\mathcal{R}, n, \gamma, \mathcal{R}(x_1), x_1),$$

where we define $\gamma = e^{-\varepsilon}$. From Theorem 7, and noting that $\beta_{m+1}(\mathcal{R}(x_1), \mathcal{R}(x_1)) = \frac{1}{m+1}$ (by Corollary 1), it follows that:

$$\begin{aligned} & \psi(\mathcal{R}, n, \gamma, \mathcal{R}(x_1), x_1) \\ &= \sum_{m=0}^{n-1} \binom{n-1}{m} \gamma^m (1-\gamma)^{n-1-m} \cdot \frac{1}{m+1} \\ &= \frac{1}{\gamma n} \sum_{m=0}^{n-1} \binom{n}{m+1} \gamma^{m+1} (1-\gamma)^{n-1-m} \\ &= \frac{1}{\gamma n} [1 - (1-\gamma)^n] \\ &\leq \frac{1}{\gamma n} = \frac{e^\varepsilon}{n}. \end{aligned}$$

This concludes the proof. $\square$

C. Analysis via the blanket decomposition

Another form of decomposition in the shuffle DP setting is the *blanket decomposition*, which extracts the *maximum* common component across all $\mathcal{R}(x_i)$ [6]. Specifically, the blanket distribution Q^B is defined pointwise as:

$$Q^B(y) := \frac{1}{\alpha} \cdot \inf_{x \in \mathbb{X}} \Pr[\mathcal{R}(x) = y], \quad \alpha := \sum_{y \in \mathbb{Y}} \inf_{x \in \mathbb{X}} \Pr[\mathcal{R}(x) = y].$$

This yields a valid convex decomposition:

$$\mathcal{R}(x) = \alpha \cdot Q^B + (1 - \alpha) \cdot \text{LO}^B(x),$$

where $\text{LO}^B(x)$ is a residual distribution specific to each input x .

The blanket decomposition is known to be optimal in deriving privacy amplification bounds under shuffle DP [14].

We now demonstrate that it is also optimal for upper bounding the success probability of re-identification attacks among all decomposition-based methods.

Theorem 9. *Let Q^B and α be the blanket distribution and its associated mass coefficient for local randomizer $\mathcal{R}$. If $\mathcal{R}$ also admits a decomposition:*

$$\forall x \in \mathbb{X}: \quad \mathcal{R}(x) = \gamma Q^{\text{com}} + (1 - \gamma) \text{LO}(x_i),$$

then

$$\beta_n(\mathcal{R}, x_1) \leq \psi(\mathcal{R}, n, \alpha, Q^B, x_1) \leq \psi(\mathcal{R}, n, \gamma, Q^{\text{com}}, x_1).$$

Proof. The first inequality follows directly from Theorem 6 by choosing $\gamma = \alpha$ and $Q^{\text{com}} = Q^B$.

To prove the second inequality, we use a hybrid argument. Since $\Pr[\mathcal{R}(x) = y] \geq \gamma Q^{\text{com}}(y)$ for all x and y by assumption, we have:

$$\inf_x \Pr[\mathcal{R}(x) = y] \geq \gamma Q^{\text{com}}(y) \Rightarrow \alpha Q^B(y) \geq \gamma Q^{\text{com}}(y).$$

This implies a decomposition of Q^B :

$$Q^B = \frac{\gamma}{\alpha} Q^{\text{com}} + \left(1 - \frac{\gamma}{\alpha}\right) Q^\ell,$$

for some distribution Q^ℓ .

In Game 5, users $i \geq 2$ sample from Q^{com} with probability γ , and from Q^ℓ with probability $\alpha - \gamma$, which together form Q^B used with probability α overall. The game also reveals the shuffled positions $\sigma^{-1}(I)$ of users who sampled from Q^ℓ , allowing the adversary to filter out those entries and behave identically to Game 3.

If this additional information is hidden, the adversary is effectively operating in the game

$$\text{GuessGame}_{\mathcal{A}}^{\mathcal{R}, n, \alpha, Q^B}(x_1).$$

Thus, any attacker for the blanket game can be simulated in Game 5 (by ignoring $\sigma^{-1}(I)$), while the converse is not true. Therefore:

$$\psi(\mathcal{R}, n, \alpha, Q^B, x_1) \leq \psi(\mathcal{R}, n, \gamma, Q^{\text{com}}, x_1). \quad \square$$

Theorem 10. *Let $\mathcal{R}$ be a local randomizer in the shuffle DP model, and fix an input $x_1 \in \mathbb{X}$. Define*

$$M := \sup_y \frac{\Pr[\mathcal{R}(x_1) = y]}{\inf_{x \in \mathbb{X}} \Pr[\mathcal{R}(x) = y]}.$$

Then the multiplicative Bayesian advantage satisfies:

$$\forall n \in \mathbb{N}, \quad \text{Adv}_n^\times(\mathcal{R}, x_1) \leq M, \quad \text{and} \quad \lim_{n \rightarrow \infty} \text{Adv}_n^\times(\mathcal{R}, x_1) = M.$$

The proof is provided in Appendix C. The proof of the upper bound follows a similar approach to that of Theorem 8. The tightness of this bound can be justified using results from the basic setting. Here, M denotes the supremum of the likelihood ratios between $\mathcal{R}(x_1)$ and all other $\mathcal{R}(x)$ for $x \in \mathbb{X}$. Roughly speaking, we may assume that this maximum is attained at $x = x^*$. Now, let $x_2 = x_3 = \dots = x_n = x^*$. This particular

Game 5: GuessGame3 $_{\mathcal{A}}^{\mathcal{R}, n, \gamma, Q^{\text{com}}}(x_1)$

```

1 Compute  $Q^B, \alpha$  of  $\mathcal{R}$ 
2  $y_1 \leftarrow \mathcal{R}(x_1) \cdot \mathbb{Y}$ 
3  $N \sim 1 + \text{Bin}(n-1, \alpha)$ 
4  $I \leftarrow \emptyset$ 
5 for  $i = 2$  to  $N$  do
6    $c_i \sim \text{Bern}(\frac{\gamma}{\alpha})$ 
7   if  $c_i = 1$  then
8      $y_i \leftarrow_{Q^{\text{com}}} \mathbb{Y}$ 
9   else
10     $y_i \leftarrow_{Q^i} \mathbb{Y}$ 
11     $I \leftarrow I \cup \{i\}$ 
12  $\mathbf{y} \leftarrow (y_1, y_2, \dots, y_N)$ 
13  $\sigma \leftarrow_{\$} \mathcal{P}_N$  // Sample a random permutation
14  $\mathbf{y}_\sigma \leftarrow (y_{\sigma(1)}, y_{\sigma(2)}, \dots, y_{\sigma(N)})$ 
15  $g \leftarrow \mathcal{A}(\mathbf{y}_\sigma, \sigma^{-1}(I))$ 
16 return  $\sigma^{-1}(1) = g$ 

```

instance of the shuffle DP setting reduces to a basic setting. By definition, we then have

$$\text{Adv}_n^\times(\mathcal{R}, x_1) \geq \beta_n(\mathcal{R}(x_1), \mathcal{R}(x^*)).$$

By Theorem 3, $\beta_n(\mathcal{R}(x_1), \mathcal{R}(x^*)) \rightarrow M$ as $n \rightarrow \infty$, and the result follows from the squeeze theorem.

On the one hand, Theorem 10 implies Theorem 8 because the differential privacy guarantee of $\mathcal{R}$ ensures that

$$\sup_y \frac{\Pr[\mathcal{R}(x_1) = y]}{\inf_x \Pr[\mathcal{R}(x) = y]} \leq e^\varepsilon.$$

This also demonstrates that the bound in Theorem 8 is tight in the worst case. On the other hand, the bound obtained from the *blanket decomposition* is input-dependent and can yield sharper bounds for specific values of x_1 . We illustrate this with the following example.

Example VI.1. Consider the Laplace mechanism defined over $[0, 1]$:

$$\mathcal{R}(x) = x + \text{Lap}\left(\frac{1}{\varepsilon}\right), \quad x \in [0, 1],$$

where $\text{Lap}(\delta)$ denotes the Laplace distribution with density function $\frac{1}{2\delta}e^{-|x|/\delta}$. A direct computation shows that the blanket distribution of $\mathcal{R}$ is $\mathcal{R}(0.5)$, and the associated mixture coefficient is $\alpha = e^{-\varepsilon/2}$.

According to Theorem 10, we have the bound

$$\text{Adv}_n^\times(\mathcal{R}, x_1) \leq \sup_y \frac{\Pr[\mathcal{R}(x_1) = y]}{\inf_x \Pr[\mathcal{R}(x) = y]} = e^{(|x_1 - 0.5| + 0.5)\varepsilon}.$$

In particular, when $x_1 = 0.5$, the bound given by the blanket decomposition yields $\text{Adv}_n^\times(\mathcal{R}, x_1) \leq e^{\varepsilon/2}$, which is strictly tighter than the clone decomposition bound $\text{Adv}_n^\times(\mathcal{R}, x_1) \leq e^\varepsilon$.

In Theorems 9 and 10, the notation $\mathcal{R}(x_i)$ is used. Here, we explain how these results naturally extend to the general case

where $x_i \sim V_i$. Following the discussion in Section III-B, it suffices to consider $\mathcal{R}^*(V_i)$. In fact, the blanket distribution of $\mathcal{R}^*$ coincides with that of $\mathcal{R}$, because all possible distributions of $\mathcal{R}^*$ are convex combinations of $\mathcal{R}$, so its maximal common component remains unchanged:

$$\forall y : \inf_{V \in \mathcal{D}[\mathbb{X}]} \Pr[\mathcal{R}^*(V) = y] = \inf_{x \in \mathbb{X}} \Pr[\mathcal{R}(x) = y].$$

Therefore, in the case where $x_i \sim V_i$, the blanket reduction remains optimal, with the only modification being to interpret $\mathcal{R}(x_1)$ as $\mathcal{R}^*(V_1)$. The interpretation for Theorem 10 is analogous.

Example VI.2. Consider $\mathcal{R}$ as a $\ln(3)$ -DP 2-RR mechanism. For $x_1 \sim V_1 = (p, 1-p)$, we have

$$\mathcal{R}(x_1) \sim (0.25 + 0.5p, 0.75 - 0.5p),$$

$$\forall y \in \{1, 2\}, \inf_{x \in \mathbb{X}} \Pr[\mathcal{R}(x) = y] = 0.25,$$

$$M = \sup_y \frac{\Pr[\mathcal{R}(x_1) = y]}{\inf_{x \in \mathbb{X}} \Pr[\mathcal{R}(x) = y]} = \max\{1 + 2p, 3 - 2p\}.$$

According to Theorem 10, when $p = 1$, we have $\text{Adv}_n^\times(\mathcal{R}, x_1) \leq 3 = e^\varepsilon$. When $p = 0.5$, $\text{Adv}_n^\times(\mathcal{R}, x_1) \leq 2 < e^\varepsilon$.

VII. CONCLUSION

In this paper, we presented the first systematic information-theoretic analysis of re-identification attacks in the shuffle model. We introduced a fundamental formulation, where one message is drawn from a distribution P and the remaining $n-1$ messages from a distribution Q , and all are anonymized via shuffling. We derived an exact analytical expression for the success probability $\beta_n(P, Q)$ of the optimal Bayesian adversary, as well as its asymptotic behavior. We further established tight mutual bounds between the Bayesian advantage and the total variation distance $\Delta(P, Q)$, highlighting their equivalence up to a $1/n$ factor.

Extending beyond this basic setting, we analyzed re-identification attacks under general shuffle differential privacy protocols. We proposed a reduction framework that transforms the shuffle DP setting into a canonical instance, enabling analysis via known quantities $\beta_n(P, Q)$. Leveraging decomposition-based techniques, we demonstrated how the *clone* and *blanket* decompositions can be used to upper bound the adversary's success probability. Notably, we showed that the blanket decomposition not only yields the optimal privacy amplification bounds in prior literature but also provides the optimal upper bound for Bayesian re-identification attacks among all decomposition-based methods.

Our results establish new theoretical foundations for understanding anonymity leakage in the shuffle model and offer new insights into the security analysis of honeyword systems, shuffle DP mechanisms, and beyond. We hope this framework can inspire further research.

REFERENCES

- [1] A. Bittau, U. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnes, and B. Seefeld, "Prochlo: Strong privacy for analytics in the crowd," in *Proceedings of the 26th Symposium on Operating Systems Principles*, ser. SOSP '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 441–459. [Online]. Available: <https://doi.org/10.1145/3132747.3132769>
- [2] A. Cheu, A. Smith, J. Ullman, D. Zeber, and M. Zhilyaev, "Distributed differential privacy via shuffling," in *Advances in Cryptology – EURO-CRYPT 2019*, Y. Ishai and V. Rijmen, Eds. Cham: Springer International Publishing, 2019, pp. 375–403.
- [3] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, "Cryptography from anonymity," in *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, 2006, pp. 239–248.
- [4] A. Beimel, I. Haitner, K. Nissim, and U. Stemmer, "On the round complexity of the shuffle model," in *Theory of Cryptography: 18th International Conference, TCC 2020, Durham, NC, USA, November 16–19, 2020, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 683–712. [Online]. Available: https://doi.org/10.1007/978-3-030-64378-2_24
- [5] U. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by shuffling: from local to central differential privacy via anonymity," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '19. USA: Society for Industrial and Applied Mathematics, 2019, p. 2468–2479.
- [6] B. Balle, J. Bell, A. Gascón, and K. Nissim, "The privacy blanket of the shuffle model," in *Advances in Cryptology – CRYPTO 2019: 39th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18–22, 2019, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, 2019, p. 638–667. [Online]. Available: https://doi.org/10.1007/978-3-030-26951-7_22
- [7] V. Feldman, A. McMillan, and K. Talwar, "Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling," in *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, 2022, pp. 954–964.
- [8] D. Wang, H. Cheng, P. Wang, J. Yan, and X. Huang, "A security analysis of honeywords," in *NDSS 2018*, 2018.
- [9] A. Juels and R. L. Rivest, "Honeywords: Making password-cracking detectable," in *ACM CCS 2013*, 2013, p. 145–160.
- [10] N. Chakraborty, J. Li, V. C. M. Leung, S. Mondal, Y. Pan, C. Luo, and M. Mukherjee, "Honeyword-based authentication techniques for protecting passwords: A survey," *ACM Comput. Surv.*, vol. 55, no. 8, dec 2022.
- [11] M. Jurado, R. G. Gonze, M. S. Alvim, and C. Palamidessi, "Analyzing the Shuffle Model Through the Lens of Quantitative Information Flow," in *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*. Los Alamitos, CA, USA: IEEE Computer Society, Jul. 2023, pp. 423–438. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/CSF57540.2023.00033>
- [12] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*. Cambridge University Press, 2022.
- [13] V. Feldman, A. McMillan, and K. Talwar, *Stronger Privacy Amplification by Shuffling for Renyi and Approximate Differential Privacy*, 2023, pp. 4966–4981. [Online]. Available: <https://epubs.siam.org/doi/abs/10.1137/1.9781611977554.ch181>
- [14] P. Su, H. Cheng, and P. Wang, "Decomposition-based optimal bounds for privacy amplification via shuffling," 2025. [Online]. Available: <https://arxiv.org/abs/2504.07414>
- [15] C. Dwork, "Differential privacy," in *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ser. ICALP'06. Berlin, Heidelberg: Springer-Verlag, 2006, p. 1–12. [Online]. Available: https://doi.org/10.1007/11787006_1
- [16] G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, and T. Wang, "Privacy at scale: Local differential privacy in practice," in *Proceedings of the 2018 International Conference on Management of Data*, ser. SIGMOD '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1655–1658. [Online]. Available: <https://doi.org/10.1145/3183713.3197390>
- [17] M. Yang, T. Guo, T. Zhu, I. Tjuawinata, J. Zhao, and K.-Y. Lam, "Local differential privacy and its applications: A comprehensive survey," *Computer Standards & Interfaces*, vol. 89, p. 103827, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0920548923001083>
- [18] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting telemetry data privately," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 3574–3583.
- [19] A. D. P. Team, "Learning with privacy at scale," *Apple Machine Learning Journal*, 2017.
- [20] [Online]. Available: <https://leak-lookup.com/>
- [21] D. Wang, Y. Zou, Q. Dong, Y. Song, and X. Huang, "How to attack and generate honeywords," in *IEEE S&P 2022*, 2022, pp. 966–983.
- [22] A. Dionysiou, V. Vassiliades, and E. Athanasopoulos, "Honeygen: Generating honeywords using representation learning," in *ACM ASIA CCS 2021*, 2021, p. 265–279.
- [23] Z. Huang, L. Bauer, and M. K. Reiter, "The impact of exposed passwords on honeyword efficacy," in *USENIX Security 2024*, 2024.
- [24] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965. [Online]. Available: <http://www.jstor.org/stable/2283137>
- [25] M. Alvim, K. Chatzikokolakis, A. McIver, C. Morgan, C. Palamidessi, and G. Smith, *The Science of Quantitative Information Flow*, ser. Information Security and Cryptography. United States: Springer, Springer Nature, 2020.
- [26] M. S. Alvim, K. Chatzikokolakis, C. Palamidessi, and G. Smith, "Measuring information leakage using generalized gain functions," in *Proceedings of the 2012 IEEE 25th Computer Security Foundations Symposium*, ser. CSF '12. USA: IEEE Computer Society, 2012, p. 265–279. [Online]. Available: <https://doi.org/10.1109/CSF.2012.26>
- [27] M. S. Alvim, K. Chatzikokolakis, A. McIver, C. Morgan, C. Palamidessi, and G. Smith, "Additive and multiplicative notions of leakage, and their capacities," in *Proceedings of the 2014 IEEE 27th Computer Security Foundations Symposium*, ser. CSF '14. USA: IEEE Computer Society, 2014, p. 308–322. [Online]. Available: <https://doi.org/10.1109/CSF.2014.29>
- [28] M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi, "On the information leakage of differentially-private mechanisms," *Journal of Computer Security*, vol. 23, no. 4, pp. 427–469, 2015. [Online]. Available: <https://doi.org/10.3233/JCS-150528>
- [29] K. Chatzikokolakis, N. Fernandes, and C. Palamidessi, "Comparing systems: Max-case refinement orders and application to differential privacy," in *2019 IEEE 32nd Computer Security Foundations Symposium (CSF)*, 2019, pp. 442–44215.
- [30] D. Wang, H. Cheng, P. Wang, X. Huang, and G. Jian, "Zipf's law in passwords," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2776–2791, 2017.
- [31] Z. Hou and D. Wang, "New observations on zipf's law in passwords," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 517–532, 2023.

APPENDIX

A. Proof of Theorem 4

Proof. Consider the case where $m-1$ of the entries in $\{y_i\}_{i=2}^n$ have the same likelihood ratio as y_1 . For the adversary to succeed, all m of these entries—including y_1 —must simultaneously attain the maximum likelihood ratio among all n inputs. Since these m entries are indistinguishable under the likelihood ranking, the attacker can do no better than guessing uniformly among them. Therefore, the success probability in this scenario is $1/m$.

We compute the total success probability:

$$\begin{aligned}
 \beta_n(P, Q) &= f(+\infty) + \sum_t \sum_{m=1}^n \frac{1}{m} f(t) \binom{n-1}{m-1} G^{n-m}(t^-) g^{m-1}(t) \\
 &= f(+\infty) + \sum_t \sum_{m=1}^n \frac{1}{n} f(t) \binom{n}{m} G^{n-m}(t^-) g^{m-1}(t)
 \end{aligned}$$

$$\begin{aligned}
&= f(+\infty) + \sum_t \sum_{m=1}^n \frac{1}{n} \cdot t \cdot g(t) \binom{n}{m} G^{n-m}(t^-) g^{m-1}(t) \\
&= f(+\infty) + \sum_t \frac{1}{n} \cdot t [(G(t^-) + g(t))^n - G^n(t^-)] \\
&= f(+\infty) + \sum_t \frac{1}{n} \cdot t [G^n(t) - G^n(t^-)],
\end{aligned}$$

where we use the identity $\frac{1}{m} \binom{n-1}{m-1} = \frac{1}{n} \binom{n}{m}$, the binomial expansion, and the identity $f(t) = t \cdot g(t)$ (Property 1). $\square$

B. Proof of Theorem 5

Proof. Lower bound. We construct an explicit adversary $\mathcal{A}^*$: it samples a random index $r \in [n]$. If $P(y_r) \geq Q(y_r)$, the adversary outputs r ; otherwise, it outputs a uniformly random index from $[n] \setminus \{r\}$.

Let σ denote the shuffling permutation. If $\sigma^{-1}(1) = r$, then $y_r \sim P$, and the attack succeeds when $P(y_r) > Q(y_r)$, which occurs with probability $\sum_{x:P(x)>Q(x)} P(x)$. Otherwise, when $\sigma^{-1}(1) \neq r$, we have $y_r \sim Q$, and the adversary wins with probability $\frac{1}{n-1}$ if $P(y_r) < Q(y_r)$, which occurs with probability $\sum_{x:P(x)<Q(x)} Q(x)$.

Thus, the total success probability is:

$$\begin{aligned}
\beta_n(P, Q) &\geq \text{GuessGame}_{\mathcal{A}^*, Q, n}^P(1) \\
&= \frac{1}{n} \sum_{x:P(x)>Q(x)} P(x) + \frac{n-1}{n} \cdot \frac{1}{n-1} \sum_{x:P(x)<Q(x)} Q(x) \\
&= \frac{1}{n} \left[\sum_{x:P(x)>Q(x)} P(x) + \sum_{x:P(x)<Q(x)} Q(x) \right] \\
&= \frac{1}{n} \sum_x \left(\frac{P(x) + Q(x)}{2} + \frac{|P(x) - Q(x)|}{2} \right) \\
&= \frac{1}{n} (1 + \Delta(P, Q)),
\end{aligned}$$

which implies:

$$\text{Adv}_n^+(P, Q) = \beta_n(P, Q) - \frac{1}{n} \geq \frac{1}{n} \cdot \Delta(P, Q).$$

Upper bound. We first introduce two supporting lemmas.

Lemma 1. Let P and Q be two probability distributions, and let $G(t)$ denote the cumulative distribution function of the likelihood ratio $\frac{P(x)}{Q(x)}$ with respect to $x \sim Q$. Define $f(+\infty) := \sum_{x:Q(x)=0} P(x)$. For $M \geq \sup_{x:Q(x)>0} \frac{P(x)}{Q(x)}$, we have

$$\int_0^M G(t) dt = M - 1 + f(+\infty).$$

Proof. We apply integration by parts and use Property 1, which states that $f(t) = t \cdot g(t)$, where g is the density of G . Note that:

$$\int_0^M G(t) dt = tG(t) \Big|_0^M - \int_0^M t dG(t)$$

$$\begin{aligned}
&= M \cdot G(M) - \int_0^M t \cdot g(t) dt \\
&= M - \int_0^M f(t) dt.
\end{aligned}$$

Since P is a probability distribution, the total mass of f is:

$$\int_0^M f(t) dt = \sum_{x:Q(x) \neq 0} P(x) = 1 - f(+\infty),$$

because $f(+\infty) = \sum_{x:Q(x)=0} P(x)$. Substituting this in yields:

$$\int_0^M G(t) dt = M - (1 - f(+\infty)) = M - 1 + f(+\infty). \quad \square$$

Lemma 2. Let P and Q be two probability distributions. Then the total variation distance between P and Q can be expressed as:

$$\Delta(P, Q) = \int_0^1 G(t) dt,$$

where $G(t)$ is defined as in Lemma 1.

Proof. We begin by writing:

$$\Delta(P, Q) = \sum_{x:P(x)<Q(x)} [Q(x) - P(x)] = \sum_{t \leq 1} [Q(S_t) - P(S_t)],$$

where $S_t := \{x \in \mathbb{X} \mid \frac{P(x)}{Q(x)} = t\}$.

Using the definition of the densities f and g over the likelihood ratio t , we can write:

$$\Delta(P, Q) = \int_0^1 [g(t) - f(t)] dt.$$

By Property 1, we know that $f(t) = t \cdot g(t)$, so:

$$g(t) - f(t) = (1 - t) \cdot g(t).$$

Thus:

$$\begin{aligned}
\Delta(P, Q) &= \int_0^1 (1 - t) \cdot g(t) dt \\
&= (1 - t)G(t) \Big|_0^1 + \int_0^1 G(t) dt \\
&= \int_0^1 G(t) dt.
\end{aligned}$$

In the last step, we used the fact that $(1 - t)G(t)$ vanishes at both $t = 0$ and $t = 1$. $\square$

We begin with the expression from Theorem 2:

$$\text{Adv}_n^+(P, Q) = \frac{1}{n} \left(M - \int_0^M G^n(t) dt \right) + f(+\infty) - \frac{1}{n}.$$

This expression holds for any $M \geq \sup_{x:Q(x)>0} \frac{P(x)}{Q(x)}$. For the derivation that follows, we require $M > 1$. To satisfy this condition, we define $M = \max \left\{ 2025, \sup_{x:Q(x)>0} \frac{P(x)}{Q(x)} \right\}$.

By Lemma 1 and Lemma 2, we have:

$$\int_0^M G(t) dt = M - 1 + f(+\infty), \quad \int_0^1 G(t) dt = \Delta(P, Q).$$

Subtracting these gives:

$$\int_1^M G(t) dt = M - 1 + f(+\infty) - \Delta(P, Q).$$

To upper bound $\text{Adv}_n^+(P, Q)$, we lower bound $\int_0^M G^n(t) dt$. Observe:

$$\int_0^M G^n(t) dt \geq \int_1^M G^n(t) dt.$$

Applying Jensen's inequality to the convex function x^n :

$$\begin{aligned} \int_1^M G^n(t) dt &\geq (M-1) \left(\frac{1}{M-1} \int_1^M G(t) dt \right)^n \\ &= (M-1) \left(1 + \frac{f(+\infty) - \Delta(P, Q)}{M-1} \right)^n. \end{aligned}$$

Substituting into the formula for $\text{Adv}_n^+(P, Q)$:

$$\begin{aligned} \text{Adv}_n^+(P, Q) &\leq \frac{1}{n} \left(M - 1 - \int_1^M G^n(t) dt \right) + f(+\infty) \\ &\leq \frac{1}{n} (M-1) \left[1 - \left(1 + \frac{f(+\infty) - \Delta(P, Q)}{M-1} \right)^n \right] + f(+\infty). \end{aligned}$$

We now apply Bernoulli's inequality:

$$(1+x)^n \geq 1+nx, \quad \text{for } x \geq -1.$$

Since $f(+\infty) - \Delta(P, Q) + M - 1 = \int_1^M G(t) dt \geq 0$, the value of $x := \frac{f(+\infty) - \Delta(P, Q)}{M-1} \geq -1$, and so:

$$\left(1 + \frac{f(+\infty) - \Delta(P, Q)}{M-1} \right)^n \geq 1 + n \cdot \frac{f(+\infty) - \Delta(P, Q)}{M-1}.$$

Substituting this into the bound:

$$\begin{aligned} \text{Adv}_n^+(P, Q) &\leq \frac{1}{n} (M-1) \left[1 - \left(1 + n \cdot \frac{f(+\infty) - \Delta(P, Q)}{M-1} \right) \right] + f(+\infty) \\ &= -(f(+\infty) - \Delta(P, Q)) + f(+\infty) = \Delta(P, Q). \end{aligned}$$

□

C. Proof of Theorem 10

Proof. Let $T := \sup_y \frac{\mathcal{R}(x_1)(y)}{Q^B(y)}$. By Theorem 7, we have:

$$\begin{aligned} \psi(\mathcal{R}, n, \alpha, Q^B, x_1) &= \sum_{m=0}^{n-1} \binom{n-1}{m} \alpha^m (1-\alpha)^{n-1-m} \cdot \beta_{m+1}(\mathcal{R}(x_1), Q^B) \\ &\leq \sum_{m=0}^{n-1} \binom{n-1}{m} \alpha^m (1-\alpha)^{n-1-m} \cdot \frac{T}{m+1} \\ &= \frac{T}{\alpha n} \sum_{m=0}^{n-1} \binom{n}{m+1} \alpha^{m+1} (1-\alpha)^{n-1-m} \end{aligned}$$

$$\begin{aligned} &= \frac{T}{\alpha n} [1 - (1-\alpha)^n] \\ &\leq \frac{T}{\alpha n}. \end{aligned}$$

Now observe that:

$$\frac{T}{\alpha} = \sup_y \frac{\mathcal{R}(x_1)(y)}{\alpha Q^B(y)} = \sup_y \frac{\Pr[\mathcal{R}(x_1) = y]}{\inf_x \Pr[\mathcal{R}(x) = y]} = M,$$

which implies the upper bound:

$$\text{Adv}_n^\times(\mathcal{R}, x_1) \leq M.$$

For the lower bound, let $x^* \in \mathbb{X}$ be such that⁴:

$$x^* = \arg \max_{x \in \mathbb{X}} \sup_y \frac{\Pr[\mathcal{R}(x_1) = y]}{\Pr[\mathcal{R}(x) = y]}.$$

Now consider the input vector where $x_2 = \dots = x_n = x^*$. This reduces the setting to the basic comparison case, so:

$$\text{Adv}_n^\times(\mathcal{R}, x_1) \geq \text{Adv}_n^\times(\mathcal{R}(x_1), \mathcal{R}(x^*)).$$

By Theorem 3, we have:

$$\text{Adv}_n^\times(\mathcal{R}(x_1), \mathcal{R}(x^*)) \rightarrow \sup_y \frac{\Pr[\mathcal{R}(x_1) = y]}{\Pr[\mathcal{R}(x^*) = y]} = M.$$

Therefore, by the squeeze theorem, we conclude that

$$\lim_{n \rightarrow \infty} \text{Adv}_n^\times(\mathcal{R}, x_1) = M.$$

This completes the proof. □

□

⁴If the arg max does not exist, then there exists a sequence (x_m^*) such that

$$\lim_{m \rightarrow \infty} \sup_y \frac{\Pr[\mathcal{R}(x_1) = y]}{\Pr[\mathcal{R}(x_m^*) = y]} = M.$$

in which case the argument follows by considering $\text{Adv}_n^\times(\mathcal{R}(x_1), \mathcal{R}(x_m^*))$.