IACR Communications in Cryptology
ISSN 3006-5496, Vol. 1, No. 4, 41 pages.

https://doi.org/10.62056/ak2i5w7sf

On the Privacy of Sublinear-Communication Jaccard Index Estimation via Min-hash

Mingyu Liang[1,2], Seung Geol Choi[a, 3], Dana Dachman-Soled[b, 1], Linsheng Liu[2] and Arkady Yerukhimovich[c, 2]

[1] University of Maryland, USA
[2] George Washington University, USA
[3] United States Naval Academy, USA

Abstract. The min-hash sketch is a well-known technique for low-communication approximation of the Jaccard index between two input sets. Moreover, there is a folklore belief that min-hash sketch-based protocols protect the privacy of the inputs. In this paper, we consider variants of private min-hash sketch based-protocols and investigate this folklore to quantify the privacy of the min-hash sketch.

We begin our investigation by presenting a highly-efficient two-party protocol for estimating the Jaccard index while ensuring differential privacy. This protocol adds Laplacian noise to the min-hash sketch counts to provide privacy protection.

Then, we aim to understand what privacy, if any, is guaranteed if the results of the min-hash are released without any additional noise, such as in the case of historical data. We begin our investigation by considering the privacy of min-hash in a centralized setting where the hash functions are chosen by the min-hash functionality and are unknown to the participants. We show that in this case the min-hash output satisfies the standard definition of differential privacy (DP) without any additional noise.

We next consider a more practical distributed setting, where the hash function must be shared among all parties and is typically public. Unfortunately, we show that in this public hash function setting, the min-hash output is no longer DP. We therefore consider the notion of *distributional differential privacy* (DDP) introduced by Bassily et al. (FOCS 2013). We show that if the honest party's set has sufficiently high min-entropy, the min-hash output achieves DDP without requiring noise.

Our findings provide guidance on how to use the min-hash sketch for private Jaccard index estimation and clarify the extent to which min-hash protocols protect input privacy, refining the common belief in their privacy guarantees.

Keywords: Differential Privacy · MPC · Sublinear Communication · Sketching · Min-hash · Jaccard Index

1 Introduction

Min-hash sketch. The min-hash sketch is a simple and well-known technique to produce an unbiased estimate of the Jaccard index [Bro97, LOZ12]. The Jaccard index [Jac01] is a similarity measure between two sets A and B, denoted $J(A, B)$, defined as the fraction of the elements in the intersection of A and B divided by the number of elements in

E-mail: mliang5@gmu.edu (Mingyu Liang), choi@usna.edu (Seung Geol Choi), danadach@umd.edu (Dana Dachman-Soled), lls@gwu.edu (Linsheng Liu), arkady@gwu.edu (Arkady Yerukhimovich)

[a] Supported in part by an NSF grant CNS-19553191.

[b] Supported in part by NSF grants CNS-2154705, CNS-1933033, and IIS-2147276.

[c] Supported in part by NSF grants CNS-2144798 (CAREER) and CNS-1955620.

Received: 2024-10-07 Accepted: 2024-12-03

their union. That is, $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$. The Jaccard index has seen wide application for clustering of websites and documents [Bro97, BGMZ97], community identification [TBK07], DNA matching [CFGT12], and machine learning [WWT$^+$19, JKWC22].

Computing the Jaccard index exactly, especially when the input sets are large, can be costly. The min-hash sketch allows communication-efficient approximation [Bro97]. The basic idea behind the min-hash sketch is to apply a random hash function h to both sets A and B and then compare the minimum hashes (denoted $\min h(A)$, $\min h(B)$) in both sets. If $\min h(A) = \min h(B)$, it means that an element in $A \cap B$ has been hashed to the minimum value among elements in $A \cup B$. This occurs with probability $J(A, B)$. Thus, to get an unbiased approximation of the Jaccard index, it suffices to repeat this procedure with sufficiently many random hashes.

Private Jaccard index via min-hash. Due to its simplicity and efficiency, the min-hash sketch has become a popular tool to approximate the Jaccard index. Moreover, since the min-hash sketch only needs to compare the minimum hashes, it has been a key building block when maintaining privacy of the input sets is important, e.g., if the input sets represent fingerprints, DNA, or medical records.

There are two classes of solutions for privacy-preserving min-hash. The first class of solutions (e.g. [CFGT12, BCG14, RCS$^+$19, Fab16]) considers how to compute the min-hash and Jaccard index in a two-party setting, where the parties do not trust each other with their private inputs. The goal of these works is to design secure two-party computation protocols for computing the min-hash sketch as efficiently as possible, but they generally do not consider the privacy implications of revealing the output. The second line of work (e.g. [YLL$^+$17, YWR$^+$19, ABS20]) considers how to make the min-hash approximation privacy-preserving by adding noise to the local min-hash sketches.

Our work. These works serve as the starting point for our study. In particular, we first present a protocol that addresses the privacy of min-hash-based approximations from two perspectives. Similar to the first class of solutions, our protocol ensures that no private information about the input is revealed beyond the Jaccard index. Additionally, in line with the second class of solutions, our protocol guarantees that even the Jaccard index output satisfies differential privacy, which is achieved through adding a small amount of noise. Next, we explore whether any variant of differential privacy can be achieved without adding noise to the protocol, which would improve its accuracy. Interestingly, we demonstrate that under specific constraints on the inputs, the resulting protocol still provides a certain level of privacy guarantees.

More formally, we define three ideal functionalities to capture flavors of min-hash. $\mathcal{F}_{\mathsf{minH}}$ computes the min-hash and then outputs *both the min-hash count and the random hashes used.* On the other hand, $\mathcal{F}_{\mathsf{privH}}$ computes the min-hash functionality and outputs *only the min-hash count.* This corresponds to a setting where the min-hash is computed by a trusted curator who does not disclose the hashes used. Finally, we define $\mathcal{F}_{\mathsf{noisy\text{-}minH}}$ which adds noise to the min-hash count computed by $\mathcal{F}_{\mathsf{minH}}$. For our first result we show that for appropriate noise levels, the $\mathcal{F}_{\mathsf{noisy\text{-}minH}}$ functionality achieves both high accuracy and differential privacy, and design a secure two-party computation of $\mathcal{F}_{\mathsf{noisy\text{-}minH}}$ that is both computation and communication-efficient. For our second result, we consider a setting in which the outputs of $\mathcal{F}_{\mathsf{minH}}$ or $\mathcal{F}_{\mathsf{privH}}$ have already been released *without added noise* and show that, under certain conditions on the inputs, this setting also provides privacy guarantees for individuals' inputs.

Differentially-private and secure computation of min-hash. To build a protocol for differentially-private min-hash we observe that the min-hash count has low global sensitivity. This allows us to define a functionality $\mathcal{F}_{\mathsf{noisy\text{-}minH}}$, parameterized by (ϵ, δ) which adds (properly-tuned) Laplace noise to the output of the min hash (See Figure 2 for details.) We then prove the following theorem about this functionality.

Theorem 1 (Informal)**.** *$\mathcal{F}_{\mathsf{noisy\text{-}minH}}$ is (ϵ, δ)-DP against an adversary corrupting either party.*

To realize a protocol for DP estimation of the Jaccard index, we now just need to instantiate this functionality. We show how this can be done efficiently using a PSI-CA functionality in Section 4. In Section 9, we evaluate the performance when instantiating the PSI-CA protocol [CGT12, TL24] in the semi-honest setting. The resulting protocol has better accuracy compared to the prior work [HTC23, ABS20]. We recommend this protocol to compute differentially-private estimates of the Jaccard index.

Privacy after leakage of min-hash output. While ideally, parties should follow recommendations to add noise to the output of the min-hash count before releasing it (as in functionality $\mathcal{F}_{\mathsf{noisy\text{-}minH}}$), in practice, this may not happen. Further, there may be historical counts that have already been released without added noise. We refer to settings in which such output is released as "output leakage." We ask whether any privacy for an individual can be salvaged in this case. Somewhat surprisingly, we show that under certain conditions on the inputs to $\mathcal{F}_{\mathsf{minH}}$ or $\mathcal{F}_{\mathsf{privH}}$, the error of the min-hash approximation itself is sufficient to achieve (variants of) differential privacy–meaning that the presence of an individual element in one of the two input sets cannot be inferred given the output of $\mathcal{F}_{\mathsf{minH}}$ or $\mathcal{F}_{\mathsf{privH}}$. Essentially, the error of the sketch acts as noise to protect the privacy of the inputs. Similar observations that sketching algorithms inherently preserve privacy under certain input restrictions have previously been shown for the Johnson-Lindenstrauss sketch [BBDS12], the LogLog sketch [CDSKY20, SSGT20], and other sketches [WPS22].

We first consider the simpler case of the privacy of an individual once the output of $\mathcal{F}_{\mathsf{privH}}$ has been released. Recall that in this setting a set of private hashes is chosen by the functionality and these hashes are not returned as output of the functionality. Standard differential privacy in this setting requires that *conditioned on knowledge of A and all but one element of B (denoted by x^*)*, the probability that the functionality outputs any value out when $x^* \in B$ versus when $x^* \notin B$ differs by a factor of at most e^ϵ with all but negligible probability.

We note that min-hash is *not* differentially private in this setting if $A \cap B$ is either too large or too small. For example, if $|A \cap B| = 0$ when $x^* \notin B$ and 1 when $x^* \in B$, then min-hash always outputs 0 in the first case and outputs a count ≥ 1 with noticeable probability in the second. We prove the following theorem showing that when this is not the case the min-hash output is differentially private:

Theorem 2 (Informal)**.** *If the size of the intersection is a constant fraction of the size of A and B, then the output of $\mathcal{F}_{\mathsf{privH}}$ is (ϵ, δ)-DP for negligible δ.*

We stress that this theorem crucially relies on the fact that the parties, and the adversary, do not have any information about the chosen hashes, and cannot learn the evaluation of the hashes on their own inputs. Note that for this theorem to be useful in a two-party protocol, the parties must compute the hashes under a 2-PC or FHE. This is unlikely to be done in practice. Thus, typically, the parties will locally store the hashes[1] during the computation. To understand the privacy of this approach, we consider the case of the $\mathcal{F}_{\mathsf{minH}}$ functionality where the output leakage includes the hash functions as well as the counts.

Unfortunately, in this case there is a problem when trying to argue privacy. In the standard DP setting, we assume that the adversary knows all of the inputs (in this case, all entries in both sets A and B) except for some input x^* and wants to determine, from the output of the computation, whether x^* was in the other party's set. If the hashes are known, then the output of min-hash is deterministic: The adversary can exactly reconstruct the min-hash execution for the case when x^* is in the set and when it is not, and then

[1] As noted previously, it is sufficient to store a short seed to identify the hash.

see which of these matches the output it received. Since the min-hash protocol provides a good approximation of the Jaccard index, the adversary will be able to exactly determine whether or not $x^* \in B$ with noticeable probability.

Note that the above attack works only if the adversary knows the entirety of both sets A, B and just tries to distinguish whether $x^* \in B$ or not. Realistically, especially when the inputs are large, the adversary would not know the entire input of the honest party. More precisely, we assume that given the adversary's set (and even the intersection between the two sets), the honest party's set still has sufficiently high min-entropy. With this assumption, we turn to the tool of distributional DP (DDP) [BGKS13] which allows us to analyze differential privacy when the distribution of inputs has sufficient uncertainty.

We begin with a relatively strong assumption on the amount of uncertainty the adversary has about the honest set. Specifically, we assume that every element that is not in the intersection is highly unpredictable (i.e., has a high amount of min-entropy), even conditioned on all the other set elements. Under this assumption, we prove the following theorem:

Theorem 3 (Informal)**.** *If each non-intersecting item has sufficiently high min-entropy, revealing the hash functions*[2] *together with the min-hash counts (as in the $\mathcal{F}_{\mathsf{minH}}$ functionality) preserves (ϵ, δ)-DDP for negligible δ, as long as the size of the intersection is a constant fraction of the size of A and B.*

Not surprisingly, the proof of this Theorem (given in Section 6) leverages the fact that when each element has individual high min-entropy, hashing each element acts as a strong randomness extractor, thus resulting in sufficient random noise for privacy.

DDP over a polynomial-size universe. However, this assumption that every item has high min-entropy is quite strong. For example, consider the setting where each item in B is chosen from a polynomial-size universe. In this case, while individual items cannot have much min-entropy, the honest party's set may still collectively have high min-entropy as long as it is large enough. Thus, for our third result, we analyze what happens under this weaker assumption that *only the full honest set, instead of each individual item, has high min-entropy.*

Note that in this case, we cannot apply the hash function as randomness extractor technique. This is because in order to guarantee that the randomness extractor yields output that is negligibly close to uniform, we must lose superlogarithmic in n bits of entropy from each input. However, in the case we are currently considering, each element has at most $O(\log n)$ bits of min-entropy. Further, we in fact have no guarantee that each element has individually high min-entropy (since the elements are not necessarily independent), but only that the total min-entropy of the non-intersection items is high. Nevertheless, we show $\mathcal{F}_{\mathsf{minH}}$ still achieves DDP, by proving a new strong chain rule for min-entropy (see Section 8.5).

Specifically, we consider the following class of distributions $\mathcal{C}$ over secret sets R of size n:

- Let $\mathcal{U}$ be a universe of polynomial size $n \cdot \ell$, where $\ell = \Omega(n^3)$.
- R is chosen uniformly from all subsets of $\mathcal{U}$ of size n.
- In general, to relax the uniformity above, we additionally allow arbitrary leakage $L = L(R)$ computed on R, such that the length of the leakage L is at most $|L| \leq c \cdot n \log \ell$, for a fixed constant $c \in (0, 1)$.
- We consider the resulting conditional distribution $\mathcal{D}$ on R given leakage L.

[2] We use cryptographic hash functions to instantiate the hashes in the random oracle model.

Theorem 4 (Informal)**.** *Assume the set R is drawn from a distribution $\mathcal{D} \in \mathcal{C}$. Then the min-hash protocol in the random oracle model (corresponding to functionality $\mathcal{F}_{\mathsf{minH}}$) preserves (ϵ, δ)-DDP for negligible δ, as long as the size of the intersection is a constant fraction of the size of A and B.*

On spoiling bits and leakage resilience. Consider a distribution over sets of n elements $R = R_1, \ldots, R_n$, where each R_i is chosen from a universe of size $\ell \in \Omega(n)$. Note that the set R can have min-entropy $\Omega(n \lg(\ell))$ while it can still be possible that for every i, the marginal distribution over R_i has only *constant* min-entropy (see Example 1.1 in [DKZ18]). To deal with such situations, Skórski [Skó19] proves a theorem showing the existence of "spoiling bits." Namely, given $R_1, \ldots, R_n$, some additional information known as spoiling bits can be released such that, conditioned on this information, for each $i \in [n]$, the distribution of R_i conditioned on $R_{<i}$, where $R_{<i}$ denotes $(R_1, \ldots, R_{i-1})$, is nearly flat (in the sense that the min/max entropy gap is at most a small additive constant). Further, the total number of spoiling bits that are released is small.

It is not hard to use Skórski's result to show that if R starts out with sufficiently high min-entropy then for a large fraction of i (those in the set $V \subseteq [n]$), the distribution of R_i conditioned on $R_{<i}$ has high min-entropy of at least $\Omega(\log(n))$, while the remaining indices (those in the set $W = [n] \setminus V$)) may have low min-entropy.

Unfortunately, this result is very brittle in the sense that the flatness conditions hold only for this particular distribution of R conditioned on the spoiled bits. Specifically, despite the flatness condition being satisfied for this distribution, the random variables R_i are *not* independent of one another. Thus, if additional information is leaked on R_j after the spoiling bits are computed, then the flatness guarantees may no longer hold for R_i.

In our setting, we require additional leakage $\{\ell_i\}_{i \in W}$ on the elements $\{R_i\}_{i \in W}$. One issue is that the set W (i.e., low min-entropy elements conditioned on the spoiling bits) is only known *after* the spoiling bits are computed. This leaves us with a dilemma:

- Leaking $\{\ell_i\}_{i \in W}$ additionally *after* the spoiling leakage can destroy the flatness property.
- On the other hand, we cannot leak $\{\ell_i\}_{i \in W}$ *before* computing the spoiling bits, since we don't know the set W yet! We could leak from all the blocks $(R_1, \ldots, R_n)$, but this may deplete the entropy needed from the random variables $\{R_i\}_{i \in V}$.

To solve this problem, we prove a new variant of the spoiling lemma that computes the spoiling bits *at the same time* as the additional leakage ℓ_i for $i \in W$ is computed so that the spoiling bits also contain $\{\ell_i\}_{i \in W}$, while still maintaining the flatness condition. The types of leakage that can be captured are essentially those such that the leakage ℓ_i for $i \in W$ can be expressed as a function of R_i and the leakages $\{\ell_j : j > i, j \in W\}$. It turns out that the leakage we need for our result has this form.

We state our theorem in general terms as we believe it may find further applications in leakage resilient cryptography. For the formal theorem statement see Theorem 5.

A note on composition. One known weakness of the DDP definition is the lack of a general composition theorem [BGKS13]. However, for the specific setting of our min-hash protocols we can leverage the small output of min-hash to argue composition properties after leakage of several outputs. Specifically, suppose that the adversary executes a min-hash protocol with (ϵ, δ)-DDP security twice with the same honest party's input both times. Since each min-hash protocol outputs a single number between 0 and k (i.e., $\lg k$ bits long), when we apply Theorem 3, the leakage profile increases to a total of at most $L + 2 \cdot \lg k$ bits. However, according to Theorem 3, as long as $|L| + 2 \lg k \leq c \cdot n \lg \ell$, each protocol execution will preserve DDP, and therefore the composition of the two protocol executions will preserve $(2\epsilon, 2\delta)$-DDP. In general, assuming that the initial leakage $|L|$ is a small constant, this type of DDP composition will hold for $O(n \cdot \frac{\lg \ell}{\lg k})$ executions.

Comparison to other approaches. We note that an alternative approach to get a differentially-private estimate of the Jaccard index is via mergeable cardinality estimation sketches (e.g. [HTC23]) to compute (an approximation of) the set intersection cardinality and use this via the inclusion-exclusion principle to compute the Jaccard index. We give a detailed comparison of error from our protocol vs. the best known cardinality estimator [HTC23] in Section 9.

2 Related Works

Differential privacy (DP). Differential privacy protects the privacy of individuals by limiting an adversary's ability to learn information about an individual input from the output of a computation [Dwo06, DMNS06]. For a good overview of differential privacy and many of the algorithms to achieve it, both in the standard curator setting and in distributed settings, we refer the reader to the book by Dwork and Roth [DR$^+$14].

Optimizing secure computation using differential privacy. Another direction of work has considered how to use DP to reduce the cost of secure computation, especially when we aim for DP-style guarantees from the final output. [BNO08] first proposed such optimization for the problem of secure summation. [HMFS17] and [GRR19] applied the differential privacy relaxation to improve efficiency of set-intersection protocols. [MG18], and [MLRG20] consider graph-parallel computations and design more efficient solutions with differential private leakages. [CCMS19] consider classic tasks like sorting, merging, and range-query data structures with differential privacy relaxation. [GKLX22] consider multiparty shuffle that allows a differentially private leakage and shows that it suffices to achieve end-to-end differential privacy in the shuffle model of DP.

Private sketching. Sketching algorithms, or "sketches" are sublinear space algorithms for approximating certain properties of large inputs or data streams. The main idea behind sketching algorithms is to generate a compact summary data structure that allows for efficient storage, merging, and processing.

Some recent works [BBDS12, CDSKY20, SSGT20, WPS22, HTC23, DTT22, LLSS19, PT22, MMNW11, MDDC16, BS15, BNSGT17, HQYC22, ZQR$^+$22] have additionally observed that sketches can often also aid in achieving privacy as the inherent loss of information in the sketch can essentially make the sketch itself be differentially private or to only require a little additional noise.

A line of research pertinent to our work involves constructing private sketches for set cardinality estimations [SNY17, STS18, NvVT20, KWS$^+$20, PS21]. Recently, [HTC23] proposed a private mergeable sketch that can be used to estimate the size of the intersection and union of sets.

Secure approximation. Secure approximation studies what functions can be securely approximated without revealing anything beyond the true output [FIM$^+$01, HKKN01]. While this notion is quite different from that of differentially private approximation that we consider here, we note that our FHE-based protocol described in Section 5 additionally achieves this.

Adversarially robust property-preserving hash functions and robust sketching. Property-preserving hash (PPH) functions allow compressing large input x into a short digest $h(x)$ such that some property $P(x, y)$ can be computed given only $h(x)$ and $h(y)$. Adversarially-robust PPH [BLV19, FS21, FLS22, HLTW22] aim to further guarantee that $P(x, y)$ is correctly computed (i.e., robust) even if the inputs x and y are chosen after the hash function h is fixed. A related concept of robust sketching, e.g. [ACSS23, BJWY22] aims to construct sketches that provide good approximations even when inputs are chosen after the randomness of the sketch is fixed.

Both of these approaches are similar to our work in that they also study the consequences of making the choice of hash (or sketch) known to the adversary. However, these works focus on robustness to adversarial inputs, while we instead focus on the privacy of the output when the adversary additionally sees the hash functions.

Differentially private min-hash DP min-hash aims to make min-hash approximation differentially private by adopting standard DP mechanisms such as adding DP noises to the output to hide individual items in the input sets. In particular, [ABS20] achieves local DP (LDP) min-hash by either adding Laplacian noise, or using generalized randomized response to perturb the minhash vectors. Other than this, there are also other earlier efforts. For example, [YWR+19] attempts to use a flawed exponential mechanism to achieve DP. This leads to a faulty claim of ϵ-DP, as pointed out in [ABS20]. [YLL+17] correctly applies exponential mechanism. However, this results in a large amount of noise being added to the results.

3 Preliminaries

A function g is *negligible*, denoted $\mathsf{negl}(\cdot)$, if for every positive integer c, there is an integer n_c such that for all $n \geq n_c$ we have $g(n) \leq 1/n^c$. Let κ denote the security parameter.

Range of hash functions and the random oracle model. We model each hash function as a random oracle that maps each item to a real value in $[0, 1]$, and the output of the hash function is long enough to ensure that the probability of any two different items having a hash collision is negligible.

Notation. Let $\mathcal{U}$ denote the universe of input elements. In this paper, we will consider two input sets $A, B \subseteq \mathcal{U}$. Let $n_A = |A|, n_B = |B|$. Let $I = A \cap B$, $n_I = |I|$. We will also let $B_{+x^*} = B \cup \{x^*\}$.

Let Eq be an equality function; i.e., $\mathsf{Eq}(a, b) = 1$ if $a = b$ and 0 otherwise. For a hash function h and a set A, we let $h(A) := \{h(a) : a \in A\}$. Let $\mathrm{B}(m, p)$ be the binomial distribution with m trials and each trial having success probability p.

Basic min-hash functionality. We describe the basic min-hash functionality in Figure 1. In this work, we will consider several variants and consider privacy implications.

The Basic Min-Hash Functionality $\mathcal{F}_{\mathsf{minH}}$

The functionality is parameterized with a random oracle $\mathcal{O}$.

Input: P_1 and P_2's input vectors $A = (x_1^A, \ldots, x_{n_A}^A)$ and $B = (x_1^B, \ldots, x_{n_B}^B)$.

Minhash:

1. Randomly sample prefix pre, which is used to define hash functions $h_1, h_2, \ldots, h_k$, where for $i \in [k]$, $h_i(\cdot) := \mathcal{O}(\mathsf{pre}||i||\cdot)$.
2. For input A, compute the min-hash vector $(u_1^A, u_2^A, \ldots, u_k^A)$ as follows:

   For each iteration $j \in [k]$:

   i. For each item $x_i^A \in A$, compute $y_{i,j}^A = h_j(x_i^A)$.
   ii. Compute the min-hash for iteration j; that is, $u_j^A = \min\{y_{i,j}^A : i \in [n_A]\}$.
3. Likewise, compute another min-hash vector $(u_1^B, u_2^B, \ldots, u_k^B)$ for input B similarly.
4. Compute $c = \sum_{j=1}^{k} \mathsf{Eq}(u_j^A, u_j^B)$.

Output: Return (pre, c) to P_1 and P_2.

Figure 1: The Basic Min-Hash Functionality

Differential privacy. The definitions of (ϵ, δ)-differential privacy are given below.

Definition 1 ((ϵ, δ)-indistinguishablity)**.** Two random variables X and Y are (ϵ, δ)-indistinguishable (denoted as $X \approx_{\epsilon,\delta} Y$) if, for all events S, we have

$$\Pr[X \in S] \leq e^{\epsilon} \cdot \Pr[Y \in S] + \delta, \quad \Pr[Y \in S] \leq e^{\epsilon} \cdot \Pr[X \in S] + \delta.$$

Definition 2 (Computational (ϵ, δ)-indistinguishablity)**.** Two random variables X and Y are computationally (ϵ, δ)-indistinguishable (denoted as $X \overset{c}{\approx}_{\epsilon,\delta} Y$) if, for any polynomial time adversary $\mathcal{A}$, it holds

$$\Pr[\mathcal{A}(X) = 1] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(Y) = 1] + \delta, \quad \Pr[\mathcal{A}(Y) = 1] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(X) = 1] + \delta.$$

Definition 3 ((Computational) (ϵ, δ)-differential privacy)**.** Let X be an input space and $\simeq_X$ be a relation capturing the notion of neighboring inputs. Let $\mathcal{M} : X \to Z$ be a randomized algorithm that takes input $x \in X$ and outputs a value over Z. We say that the mechanism $\mathcal{M}$ is (ϵ, δ)-differentially private if the following holds:

$$\forall x, x' \in X \text{ s.t. } x \simeq_X x' : \ \mathcal{M}(x) \approx_{\epsilon,\delta} \mathcal{M}(x').$$

The mechanism $\mathcal{M}$ is (ϵ, δ)-computationally differentially private if $\forall x, x' \in X$ s.t. $x \simeq_X x' : \ \mathcal{M}(x) \overset{c}{\approx}_{\epsilon,\delta} \mathcal{M}(x')$.

Definition 4. The global sensitivity of a function $f : \mathbf{N}^{|\mathcal{X}|} \to \mathbf{R}^k$ is:

$$\Delta f = \max_{X,Y \in \mathbf{N}^{|\mathcal{X}|}, \|X-Y\|_1 = 1} \|f(X) - f(Y)\|_1$$

Definition 5. The Laplace Distribution (centered at 0) with scale b is the distribution with probability density function: $Lap(x|b) = \frac{1}{2b} e^{-|x|/b}$.

We will write $Lap(b)$ to denote the Laplace distribution with scale b. Given any function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$, the Laplace mechanism that adds noise drawn from Laplace distribution; that is, given an input database X, the mechanism outputs $f(X) + (Y_1, \ldots, Y_k)$, where Y_i are i.i.d. random variables drawn from $Lap(\Delta f/\epsilon)$. It is known that the Laplace mechanism achieves $(\epsilon, 0)$-differential privacy [DR+14, Theorem 3.6].

Distributional differential privacy (DDP). We adapt the original definition [BGKS13] for our purpose to consider a two-party protocol that takes sets as input more explicitly. Specifically, we consider a computational indistinguishability variant for our DDP definition.

Definition 6 (View of a party in a two-party functionality)**.** Given a two-party functionality $\mathcal{F}$ with parties P_1 and P_2, let $\mathsf{view}^{\mathcal{F}}_{P_1}(A, B)$ denote the view of P_1 for the execution of functionality $\mathcal{F}$ with A and B being the input of P_1 and P_2 respectively. In particular, $\mathsf{view}^{\mathcal{F}}_{P_1}(A, B)$ consists of the following (the view of P_2 is defined similarly):

- The input A of P_1, the private random coins of P_1, and the output of the functionality.
- If the functionality is in the random oracle model, we allow a semi-honest P_1 to make a polynomial number of arbitrary queries to the random oracle and to add the input/output information to its view.

Definition 7 (DP and DDP of a two-party functionality)**.** A two party functionality $\mathcal{F}$ is (computationally) (ϵ, δ)**-DP** against an adversary corrupting P_1, if for every (A, B) and every $x^* \in \mathcal{U}$, it holds that $\mathsf{view}^{\mathcal{F}}_{P_1}(A, B)$ is (compuatationally) (ϵ, δ)-indistinguishable from $\mathsf{view}^{\mathcal{F}}_{P_1}(A, B_{+x^*})$.

Let $\mathcal{X}$ denote a random variable for two sets over universe $\mathcal{U}$. Let $\mathcal{Z}$ denote the random variable measuring the additional auxiliary information known to the adversary.

A two party functionality $\mathcal{F}$ is (computationally) $(\epsilon, \delta, \Delta)$**-DDP** against an adversary corrupting P_1, if for every distribution $\mathcal{D} \in \Delta$ on $(\mathcal{X}, \mathcal{Z})$, every $(X = (A, B), Z)$ in the support of $(\mathcal{X}, \mathcal{Z})$ and every $x^* \in \mathcal{U}$, it holds that $\left(\mathsf{view}^{\mathcal{F}}_{P_1}(A, B), Z\right)$ is (computationally) (ϵ, δ)-indistinguishable from $\left(\mathsf{view}^{\mathcal{F}}_{P_1}(A, B_{+x^*}), Z\right)$. Here, (A, B) and Z are sampled from $\mathcal{D}$, and each party may use additional randomness.

DP and DDP against an adversary corrupting P_2 is defined symmetrically.

Tail bound for a Binomial distribution. We will use this well-known inequality.

Lemma 1 ([Doe18])**.** *Consider a Binomial distribution $B(n, p)$. We have*

$$\Pr_{X \sim B(n,p)}[X \geq k] \leq \binom{n}{k} p^k.$$

4 Min-Hash with DP

Since the noiseless min-hash functionality cannot achieve DP as discussed above, we consider a noisy variant that provides DP. We first consider the global sensitivity of $\mathcal{F}_{\mathsf{minH}}$ and use the standard Laplace mechanism to provide DP.

4.1 Sensitivity

Let $B = (x^B_1, \ldots, x^B_{n_B})$ and $B_{+x^*} = (x^B_1, \ldots, x^B_{n_B}, x^*)$, and WLOG, we consider two neighboring inputs (A, B) and (A, B_{+x^*}); the case in which x^* is added into A can be shown symmetrically.

We show how changing the input sets from B to B_{+x^*} affects the final count. Let x^* be the $(n_B + 1)$-th element of B_{+x^*}. Consider iteration j of Step 2 in Figure 1. Since we model each hash function h_j as a random oracle, $(y^B_{1,j}, \ldots, y^B_{n_B+1,j})$ will be uniformly distributed. Now, consider how the min-hash u^B_j is computed. *The value x^* from B_{+x^*} can affect the min-hash u^B_j (and thereby the final count c), only if $y^B_{n_B+1,j}$ is smaller than $(y^B_{1,j}, \ldots, y^B_{n_B,j})$.*

The probability that $y^B_{n_B+1,j}$ will be less than all $y^B_{i,j}$s is at most $1/(n_B + 1)$ by a symmetry argument. Note the final output is computed as the sum of k of these trials. Let

$$S_{x^*} = \left\{ j \in [k] : y^B_{n_B+1,j} < \min_{i \in [n_B]} \{y^B_{i,j}\} \right\}.$$

Therefore, we consider a binomial distribution $|S_{x^*}| \sim \mathrm{B}(k, 1/(n_B + 1))$, which represents how many iterations j cause x^* to be the min-hash u^B_j. In other words, $|S_{x^*}|$ captures the sensitivity of min-hash. Therefore, given the failure probability δ, the following measure can be used as the global sensitivity:

$$\sigma(\delta, k, n_B) := \arg\min_s \ \{s : \Pr_{h_1,\ldots,h_k}[|S_x| \geq s] \leq \delta\}$$

Lemma 2. *For any $\{x^B_i\}_{i \in [n_B]}$ and $x \in \mathcal{U}$, we have $\sigma(\delta, k, n_B) \leq \binom{k}{s} \cdot \left(\frac{1}{n_B+1}\right)^s$.*

Proof. The result immediately follows from Lemma 1. □

According to the above lemma, Asymptotically, with $k = \Omega(\kappa)$, we have $\sigma(\delta = \mathsf{negl}(\kappa), k, n = \Theta(k^2)) = O(\lg \lg k)$.

Noisy Min-Hash Functionality $\mathcal{F}_{\textsf{noisy-minH}}$

1. Run $\mathcal{F}_{\textsf{minH}}$ and let $(\textsf{pre}, c)$ be the output from $\mathcal{F}_{\textsf{minH}}$.
2. Sample ζ^A and ζ^B from Laplace distributions $Lap(\frac{\sigma(\delta/2,k,n_A)}{\epsilon})$ and $Lap(\frac{\sigma(\delta/2,k,n_B)}{\epsilon})$ respectively. Let $\textsf{coins}_A$ and $\textsf{coins}_B$ the random coins that the functionality used to sample them.
3. Let ℓ_A and ℓ_B be the minimum integers satisfying $\Pr[|\zeta^A| > \ell_A] \le \delta/2$ and $\Pr[|\zeta^B| > \ell_B] \le \delta/2$. If $|\zeta_A| > \ell_A$, then truncate it so that it holds $|\zeta_A| = \ell_A$. Likewise, truncate ζ_B if necessary.
4. Let $\textsf{r}(\cdot)$ be a rounding function. Output $(\textsf{pre}, \textsf{coins}_A, \textsf{r}(c + \zeta^B))$ to P_1 and $(\textsf{pre}, \textsf{coins}_B, \textsf{r}(c + \zeta^A))$ to P_2.

Figure 2: Noisy Min-Hash Functionality

4.2 Noisy Min-Hash

We consider a variant $\mathcal{F}_{\textsf{noisy-minH}}$ of $\mathcal{F}_{\textsf{minH}}$ described in Figure 2.

Theorem 1. *$\mathcal{F}_{\textsf{noisy-minH}}$ is (ϵ, δ)-DP against an adversary corrupting either party.*

Proof. Based on the definition, $\sigma(\cdot)$ works as the upperbound on the sensitivity with probability $1 - \delta/2$. For the honest party's noise (i.e., ζ_A or ζ_B), truncation takes place with probability at most $\delta/2$. Therefore, using the standard Laplace mechanism [DR+14, Theorem 3.6], and since DP is preserved even with post-processing, $\mathcal{F}_{\textsf{noisy-minH}}$ provides (ϵ, δ)-DP. □

A two party protocol $\pi_{\textsf{NMH}}$ securely realizing $\mathcal{F}_{\textsf{noisy-minH}}$. We construct a two party protocol that securely realizes functionality $\mathcal{F}_{\textsf{noisy-minH}}$. The protocol takes advantage of an ideal functionality $\mathcal{F}_{\textsf{psi-ca}}$ of private set intersection cardinality (PSI-CA) [CGT12] that computes the exact cardinality of the intersection of the two input sets, as described in Figure 3.

Functionality of Private Set Intersection Cardinality $\mathcal{F}_{\textsf{psi-ca}}$

Input: P_1 has a set A and P_2 has a set B.
Output: Return $|A \cap B|$ to P_1 and $\perp$ to P_2.

Figure 3: Functionality $\mathcal{F}_{\textsf{psi-ca}}$

In particular, in order to compute the noisy min-hash match counts, the parties construct two sets consisting min-hash values and additional dummy elements and then run $\mathcal{F}_{\textsf{psi-ca}}$ on these sets. To reflect the Laplace noise into elements of a set, the protocol uses unary encoding, which introduces some inefficiency. However, as the tail probability of Laplace noise decreases exponentially, the unary encoding length can be bounded with a small value, and the protocol's overall efficiency is still maintained. Detailed steps of the protocol are provided in Figure 4.

It is worth noting that the above task could also be implemented using a generic two-party computation (2PC) protocol. However, [CGT12] proposed an efficient PSI-CA protocol that outperforms 2PC protocols for small input sizes (using the start-to-finish comparison including the 2PC preprocessing steps). See Section 9.1 for more details of this PSI-CA protocol. Since our input sets are small, we chose to present protocol $\pi_{\textsf{NMH}}$ using the PSI-CA functionality.

We will prove below that protocol $\pi_{\textsf{NMH}}$ securely realizes $\mathcal{F}_{\textsf{noisy-minH}}$. It implies that protocol $\pi_{\textsf{NMH}}$ is also (ϵ, δ)-computational-DP [SCRS17]. The main benefit of the protocol is that the hash computations can be computed locally and the communication complexity of the protocol is *sub-linear in n_A and n_B* even when the protocol implementing $\mathcal{F}_{\textsf{psi-ca}}$ has a linear communication complexity.

Two-party Noisy Min-hash Protocol $\pi^{\mathcal{O}}_{\textsf{NMH}}$

Input: P_1 and P_2's input vectors $A = (x^A_1, \dots, x^A_{n_A})$ and $B = (x^B_1, \dots, x^B_{n_B})$.

Protocol:

1. P_1 samples prefix $\textsf{pre}$ and sends it to P_2. This prefix is used to define hash functions $h_1, h_2, \dots, h_k$, where for $i \in [k]$, $h_i(\cdot) := \mathcal{O}(\textsf{pre}||i||\cdot)$.
2. P_1 computes the min-hash vector $(u^A_1, u^A_2, \dots, u^A_k)$ locally exactly as described in $\mathcal{F}_{\textsf{minH}}$. Likewise, P_2 locally computes $(u^B_1, u^B_2, \dots, u^B_k)$.
3. P_1 (resp. P_2) samples ζ^A (resp. ζ^B) from Laplace distribution $Lap(\frac{\sigma(\delta/2,k,n_A)}{\epsilon})$ (resp. $Lap(\frac{\sigma(\delta/2,k,n_B)}{\epsilon})$). Let $\textsf{coins}_A$ (resp. $\textsf{coins}_B$) be the random coins that P_1 (resp. P_2) used in sampling the noise ζ^A (resp. ζ^B). As in $\mathcal{F}_{\textsf{noisy-minH}}$, parties truncate ζ^A and ζ_B based on ℓ_A and ℓ_B, if necessary.

   Let Z^B be a $2\ell_B$-bit vector representing the unary encoding of $\mathsf{r}(\zeta^B + \ell_B)$. That is, the first $\mathsf{r}(\zeta^B + \ell_B)$ bits are 1's and the remaining bits are 0's. We let Z^B_j denote the jth bit of Z^B.
4. P_1 and P_2 invokes $\mathcal{F}_{\textsf{psi-ca}}$ with the following inputs:
   - P_1's input: $\{(i, u^A_i) : i \in [k]\} \cup \{(j + k, 1) : j \in [2\ell_B]\}$
   - P_2's input: $\{(i, u^B_i) : i \in [k]\} \cup \{(j + k, Z^B_j) : j \in [2\ell_B]\}$

   Let *out* be the output to P_1 from functionality $\mathcal{F}_{\textsf{psi-ca}}$. Set $c_A = out - \ell_B$.
5. P_1 computes $c^+ = c^A + \mathsf{r}(\zeta^A)$ and sends c^+ to P_2. P_2 computes $c^B = c^+ - \mathsf{r}(\zeta^B)$.

Output: P_1 and P_2 output $(\textsf{pre}, \textsf{coins}_A, c^A)$ and $(\textsf{pre}, \textsf{coins}_B, c^B)$ respectively.

Figure 4: A two-party min-hash protocol with noise

Proposition 1. *Protocol $\pi^{\mathcal{O}}_{\textsf{NMH}}$ described in Figure 4 securely realizes $\mathcal{F}_{\textsf{noisy-minH}}$ in the semi-honest model.*

Proof. First note that the protocol will correctly compute $c_A = \mathsf{r}(c+\zeta^B)$ and $c_B = \mathsf{r}(c+\zeta^A)$ as in $\mathcal{F}_{\textsf{noisy-minH}}$. For privacy, when P_1 is corrupted, the only message to simulate is c_A, the output from $\mathcal{F}_{\textsf{psi-ca}}$. Since the protocol is in the $\mathcal{F}_{\textsf{psi-ca}}$ hybrid, this message c^A can be perfectly simulated by using the output from $\mathcal{F}_{\textsf{noisy-minH}}$. The simulator can also make sure that $\textsf{pre}$ and ζ^A are correctly sampled by using $\textsf{pre}$ and $\textsf{coins}_A$ from $\mathcal{F}_{\textsf{noisy-minH}}$. For corrupted P_2, first the simulator makes sure that $\textsf{pre}$ and ζ^B are correct by using $\textsf{pre}$ and $\textsf{coins}_B$ from $\mathcal{F}_{\textsf{noisy-minH}}$. The message $c^+ = c^B + \mathsf{r}(\zeta^B)$ can also be perfectly simulated, since the simulator can obtain c^B from $\mathcal{F}_{\textsf{noisy-minH}}$. □

5 Noiseless Protocol in the Private Hash Setting

In Figure 5, we describe the min-hash protocol $\mathcal{F}_{\textsf{privH}}$ in the private hash setting. We show that if $J(A, B)$ is a constant, there exist parameter regimes where $\mathcal{F}_{\textsf{privH}}$ without noise satisfies differential privacy. Our observation is that the final count c follows a binomial distribution in the private hash setting, which can be treated as noise to obscure the sensitivity.

Min-Hash in the Private Hash Setting $\mathcal{F}_{\mathsf{privH}}$

$\mathcal{F}_{\mathsf{privH}}$ works exactly the same as $\mathcal{F}_{\mathsf{minH}}$ except that it outputs only the final count c (with the prefix pre hidden to the participants).

Figure 5: Min-Hash in the Private Hash Setting

Theorem 2. *For any constant $\epsilon > 0$, if $k = k(\epsilon, \kappa) \in \Omega(\kappa), n_A/k \in \Omega(\kappa), n_B/k \in \Omega(\kappa)$, and $J(A, B) \in (0, 1)$ is a constant independent of κ, then $\mathcal{F}_{\mathsf{privH}}$ is (ϵ, δ)-DP with $\delta \in \mathsf{negl}(\kappa)$.*

Proof. WLOG, let $B = (x_1^B, \dots, x_{n_B}^B)$ and $B_{+x^*} = (x_1^B, \dots, x_{n_B}^B, x^*)$. Let $p = J(A, B)$, $s = \sigma(\delta, k, \min(n_A, n_B)) = O(\lg\lg\kappa)$. Recall the definition S_{x^*} in Section 4.1 and let $K_{x^*} = [k] \setminus S_{x^*}$. Note that for the iterations in K_{x^*}, the min-hash matches (denoted as $c_{K_{x^*}}$) for both (A, B) and (A, B_{+x^*}) will be identically distributed. This match count c_{x^*} will work as an additive noise. Since $h_1, \dots, h_k$ are private, we have $c_{K_{x^*}} \sim \mathrm{B}(k - s, p)$. By applying Lemma 3 below, we conclude that $\mathcal{F}_{\mathsf{privH}}$ is differentially private. □

Lemma 3. *Consider a Binomial distribution $B(n, p)$, where $n \in \Omega(\kappa)$ and $p \in (0, 1)$ is a constant independent of κ. Then, for any constant ϵ and $s = O(\lg\lg\kappa)$, there are $a, b \in [n]$ with $a < np < b$ such that*

- *For any $\ell \in [a, b], e^{-\epsilon} \leq \frac{\Pr_{X\sim B(n,p)}[X=\ell]}{\Pr_{X\sim B(n,p)}[X+s=\ell]} \leq e^{\epsilon}$.*
- *For any $\ell \notin [a, b]$, $\Pr[B(n, p) = \ell] = \mathsf{negl}(\kappa)$ and $\Pr[B(n, p) + s = \ell] = \mathsf{negl}(\kappa)$.*

The proof of the lemma is found in Appendix A.

Remark. While $\mathcal{F}_{\mathsf{privH}}$ could be considered as a trusted curator model, a two-party protocol realizing it can be constructed without relying on a trusted curator. In particular, the computation of $(u_1^A, \dots, u_k^A)$ (including all n hash evaluations) can be performed locally under a threshold FHE so that only the encryption of them may be sent to party B. Then, by computing the remaining steps under FHE and delivering the result using a threshold decryption, the protocol will securely realize $\mathcal{F}_{\mathsf{privH}}$ in the semi-honest setting. We note that the resulting protocol has sublinear communication in n since only the k inputs to the comparisons need to be communicated.

6 DDP of $\mathcal{F}_{\mathsf{minH}}$

In this section, we show that there are parameter regimes where the public min-hash protocol $\mathcal{F}_{\mathsf{minH}}$ can satisfy DDP without adding noise. In Figure 6, we first describe the family of distributions we consider in the context of our min-hash protocol. The distribution models a situation in which the adversary, having corrupted one of the two parties, has access to the view of the party and even the actual intersection. However, the adversary does not know the other party's input set (except from the intersection).

Below, we show that $\mathcal{F}_{\mathsf{minH}}$ achieves DDP under certain circumstances.

Theorem 3. *For every constant $\epsilon > 0$, consider $\mathcal{F}_{\mathsf{minH}}$ in the random oracle model with $k = k(\epsilon, \kappa)$, where $k \in \Omega(\kappa)$. Let $R = B \setminus I$, each element of which has min-entropy at least κ. Let $n_A/k, n_B/k \in \Omega(\kappa)$, and $n_I/n_A \in (0, 1)$ is a constant independent of κ. Then, $\mathcal{F}_{\mathsf{minH}}$ is computationally $(\epsilon, \delta, \Delta_{\mathsf{PH}})$-DDP against an adversary corrupting P_1 with $\delta \in \mathsf{negl}(\kappa)$. DDP against an adversary corrupting P_2 holds when the parameters are set symmetrically.*

Theorem 4. *For security parameter κ, every constant $\epsilon > 0$, and every constant $\gamma \in (0, 1)$, consider $\mathcal{F}_{\mathsf{minH}}$ in the random oracle model with $k = k(\epsilon, \kappa)$, where $k \in \Omega(\kappa \cdot \lg\lg\kappa)$. Let*

Distribution Family Δ_{PH}

Parameterized with (n_A, n_B, n_I), a distribution $\mathcal{D}_{A,B}$ in this family samples (A, B) such that

- Letting $I = A \cap B$, it holds that $|A| = n_A$, $|B| = n_B$, and $|I| = n_I$

Output:

- The inputs to the parties P_1 and P_2 are A and B respectively.
- Give I to the adversary as the auxiliary information.

Figure 6: The family of distributions that we consider in our min-hash protocol

$R = B \setminus I$ be a set of size n_R, with $n_R/k^2 \in \Omega(\kappa)$. Let the universe $\mathcal{U}$ be of size $n_R \cdot \ell$, where $\ell = \Omega(n_R^3)$. Assume the secret set R is chosen chosen uniformly from all subsets of $\mathcal{U}$ of size n_R, conditioned on arbitrary leakage on R of length L, where $n_R \lg \ell - L \geq \frac{8n_R}{9} \lg \ell + 2n_R$. Let $|I| \in \Theta(n)$. Then the output of $\mathcal{F}_{\mathsf{minH}}$ achieves computational $(\epsilon, \delta, \Delta_{\mathsf{PH}})$-DDP with $\delta \in \mathsf{negl}(\kappa)$ against an adversary corrupting P_1. DDP against an adversary corrupting P_2 holds when the parameters are set symmetrically.

Remark. An easy way to realize $\mathcal{F}_{\mathsf{minH}}$ is to have each party locally hash their inputs using the k public hash functions and to locally compute the minimum for each iteration. The parties can then run a simple two-party computation to compute the number of times these minimums match. We note that this protocol has communication and computation that is sublinear in the input size as it only depends on the number of hash functions. By Theorems 3 and 4 this protocol achieves DDP when the conditions of either of the theorems are satisfied.

7 Proof of Theorem 3

We first give the intuition of the proof. We assume that each of the non-intersecting elements has high min-entropy. WLOG, consider an adversary corrupting P_1. The view of the adversary will be

$$\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{minH}}}(A, B) := (A, c, h_1, \ldots, h_k).$$

As shown above, the sensitivity can be upper-bounded by a small value s.

Unlike $\mathcal{F}_{\mathsf{privH}}$, however, when we show the existence of sufficient noise from the remaining iterations, we need to take the additional leakage into consideration.

First, since the hash functions are public, iterations are no longer independent of each other as needed by the analysis in Section 5. We address this issue by employing the fact that each of the non-intersecting items has high min-entropy. In the random oracle model, as long as the adversary does not query hash function h on some point x, $h(x)$ is uniformly random to the adversary. Since the non-intersecting items have high min-entropy, the adversary is negligibly likely to query any of them to the hash functions, thus guaranteeing independence.

Good iterations and Poisson Binomial distribution. Now, to see how the remaining iterations still hide the sensitivity even with the public hash functions, let $R = B \setminus I$. For the remaining $k - s$ iterations, the high min-entropy of each element in R will jitter the final count. In particular, consider the jth hash function h_j in the protocol (among the $k - s$ remaining iterations) and let

$$v_j^A = \min h_j(A), \quad v_j^I = \min h_j(I), \quad v_j^R = \min h_j(R).$$

Suppose $v_j^A = v_j^I$. Then, if $v_j^R \geq v_j^I$, the min-hash u_j^A of A will be equal to the min-hash u_j^B of B (both of which are equal to v_j^I) and the final count c will be incremented due to this jth iteration. However, if $v_j^R < v_j^I$, then it will be $u_j^A \neq u_j^B$, and the final count will not be incremented. This way, the distribution of v_j^R will jitter the final count. The above discussion can be formalized into the following definition.

Definition 8 (θ-good iteration)**.** Let $n_R = n_B - n_I$, we define $\mathsf{good}_\theta(h_j, A, I, n_B)$ to be true if and only if the following holds:

$$\min h_j(A) = \min h_j(I), \text{ and } \min h_j(I) \in \left[1 - \left(\frac{1}{2} + \theta\right)^{1/n_R}, 1 - \left(\frac{1}{2} - \theta\right)^{1/n_R}\right].$$

The second condition of the definition requires that $\min h_j(I)$ is somewhere in the middle (parameterized by $\theta \in \Theta(1)$) so that the distribution of R (i.e., random v_j^R) may reduce the final count with a decent chance (and also keep the count with a decent chance). As long as n_I/n_A is a constant fraction, there are sufficiently many θ-good iterations, although we lose some iterations. In particular, if we let k_g be the number of good iterations, we have $k_g = \Theta(k)$.

With public hash functions and thereby $\min h_j(I)$ being leaked to the adversary, it turns out that the noise from the k_g iterations follows a Poisson Binomial distribution, which is a generalization of a Binomial distribution where each trial has a different success probability. However, using the techniques of [COK22], we can still show that this distribution works as a good noise to hide the private data.

7.1 Proof

WLOG, we consider two neighboring inputs (A, B) and (A, B_{+x^*}). DDP for the case in which x^* is added into A can be shown symmetrically. We prove the theorem by a hybrid argument. In particular, we define a slightly different ideal functionality $\mathcal{F}_{\mathsf{minH}}{}^{(1)}$ as follows:

- Let $\mathcal{F}_{\mathsf{minH}}{}^{(1)}$ be the same as $\mathcal{F}_{\mathsf{minH}}$ except that for each $x_i^B \in B \setminus A$, each element in $\{y_{i,j}^B\}_j$ is chosen uniformly at random from $[0, 1]$.

We set up the following hybrids. We will argue that for any $x^* \in \mathcal{U}$ and over $(A, B, I) \leftarrow \Delta_{\mathsf{PH}}$, it holds

$$\begin{aligned}(\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{minH}}}(A,B), I) &\overset{c}{\approx} (\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{minH}}{}^{(1)}}(A,B), I)\\ &\approx_{\epsilon,\delta} (\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{minH}}{}^{(1)}}(A,B_{+x^*}), I) \overset{c}{\approx} (\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{minH}}}(A,B_{+x^*}), I)\end{aligned}$$

for any constant $\epsilon > 0$ and for some $\delta \in \mathsf{negl}(\kappa)$, as long as each element in $B \setminus I$ has high min-entropy.

Recall that the min-entropy of each element x_i^B with $i \in B \setminus A$ is at least κ. Therefore, the probability that any adversary making at most polynomially many oracle queries queries any x_i^B is $\mathsf{negl}(\kappa)$. Conditioned on the adversary not querying any such x_i^B, any $y_{i,j}^B$ for $j \in [k]$ is chosen uniformly random from $\mathcal{U}$. The same argument shows $\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{minH}}}(A, B_{+x^*}), I) \overset{c}{\approx} (\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{minH}}{}^{(1)}}(A, B_{+x^*}), I)$. Therefore, we are left only to show $(\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{minH}}{}^{(1)}}(A, B), I) \approx_{\epsilon,\delta} (\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{minH}}{}^{(1)}}(A, B_{+x^*}), I)$.

DDP of $\mathcal{F}_{\mathsf{minH}}{}^{(1)}$. We show $(A, I, h_1, \ldots, h_k, c) \approx_{\epsilon,\delta} (A, I, h_1, \ldots, h_k, c_{+x^*})$, where c is the final count from $\mathcal{F}_{\mathsf{minH}}{}^{(1)}(A, B)$ and c_{+x^*} is the final count from $\mathcal{F}_{\mathsf{minH}}{}^{(1)}(A, B_{+x^*})$. We show how to leverage the uncertainties of $x_i^B \in R = B \setminus A$ so that good iterations work like the needed noise to guarantee DP.

Lemma 4. *For any A, I, n_B and $n_R = n_B - n_I$, we have*

$$p_\theta \overset{def}{=} \Pr_h[\mathsf{good}_\theta(h, A, I, n_B)] \geq \left(\left(\frac{1}{2} + \theta\right)^{\frac{n_A}{n_R}} - \left(\frac{1}{2} - \theta\right)^{\frac{n_A}{n_R}}\right) \cdot \frac{n_I}{n_A} \quad .$$

The proof is found in Appendix B. This lemma shows that a random hash leads to a good iteration with probability p_θ, which is constant in our setting based on the assumption about n_A, n_I, n_R.

Recall that S_{x^*} was the random variable that represents the set of iterations j such that the min-hash u_j^B comes from x^* when P_2's input is B_{+x^*}. From Lemma 2, with overwhelming probability $|S_{x^*}| = O(\lg\lg\kappa)$.

Now, fix A, I, x^* and $h_1, \ldots, h_k$ and let G_θ be the set of iterations j in which a θ-good event takes place; i.e.,

$$G_\theta = \{j \in [k] : \mathsf{good}_\theta(h_j, A, I, n_B)\}\,.$$

Let $K_\theta = G_\theta \setminus S_{x^*}$. The following lemma shows that the θ-good events takes place $\Theta(\kappa)$-many times, with overwhelming probability.

Lemma 5. *Suppose $k = \Theta(\kappa)$, $n_B = \Omega(\kappa^2)$, and $p_\theta \in \Theta(1)$. Let $s = |S_{x^*}|$. Then, we have*

$$\Pr_{h_1,\ldots,h_k}\left[|K_\theta| > \frac{2}{3}(k-s)p_\theta\right] \geq 1 - \mathsf{negl}(\kappa).$$

The proof is found in Appendix C.

Our goal. For a set W, define $c_W \overset{def}{=} \sum_{j\in W} \mathsf{Eq}(u_j^A, u_j^B)$. Let $\overline{K}_\theta := [k] \setminus K_\theta$. Note that the contributions to the final output can be divided into two parts:

- c_{K_θ}: The contribution from the iterations in K_θ, which contains all the θ-good iterations such that x^* does not hash to the minimum across B_{+x^*}.
- $c_{\overline{K_\theta}}$: The contribution from all the remaining iterations

Essentially, for any final count q, we are interested in comparing the two probabilities:

$$\Pr[c_{\overline{K}_\theta} + c_{K_\theta} = q] \text{ and } \Pr[c_{\overline{K}_\theta}^{+x^*} + c_{K_\theta}^{+x^*} = q].$$

Following our discussion on sensitivity in Section 4, the difference of $c_{\overline{K}_\theta}$ and $c_{\overline{K}_\theta}^{+x^*}$ is upper-bounded by $s = O(\lg\lg\kappa)$. Note that we have $c_{K_\theta} = c_{K_\theta}^{+x^*}$ because $j \in K_\theta$ implies $j \notin S_{x^*}$. Therefore, we only need to analyze the single distribution of c_{K_θ} as a noise and compare the following two probabilities:

$$\Pr[c_{K_\theta} = q] \text{ and } \Pr[c_{K_\theta} + s = q].$$

Distribution of c_{K_θ}. We have $c_{K_\theta} = \sum_{j\in K_\theta} c_j$, where $c_j = \mathsf{Eq}(u_j^A, u_j^B)$. Note that since we have $j \in K_\theta$, a θ-good event takes place in iteration j, i.e., $\min h_j(A) = \min h_j(I)$.

Let $\gamma_j = 1 - \min h_j(I)$. Note that the hash of each item R is randomly distributed in $\mathcal{F}_{\mathsf{minH}}{}^{(1)}$. Therefore, the probability that $c_j = 1$ is $(\gamma_j)^{n_R}$, in which case every hash of items in R must be at least $\min h_j(I)$.

Let $\eta_{-\theta} = 1/2 - \theta$ and $\eta_{+\theta} = 1/2 + \theta$. Since j is a good iteration, we have $(\gamma_j)^{n_R} \in [\eta_{-\theta}, \eta_{+\theta}]$. Therefore, letting $p_j = (\gamma_j)^{n_R}$, we have $c_j \sim \textsc{Ber}(p_j)$, where $\textsc{Ber}$ denotes the Bernoulli distribution. Since these Bernoulli distributions are independent from each other, can apply Lemma 6 below to conclude that $C_{K_\theta} \approx_{\epsilon,\delta} C_{K_\theta} + s$.

For $j \in [n]$, consider $c_j \sim \textsc{Ber}(p_j)$. With $p_J = \{p_j\}_{j=1}^n$, let $\mathsf{PB}(n, p_J)$ denote the distribution of $\sum_{j\in[n]} c_j$. This distribution is called a Additive Poisson Binomial distribution.

We conclude the proof by showing that the Additive Poisson Binomial distribution with appropriate parameters satisfies the following DP-like property.

Lemma 6. *Consider an Additive Poisson Binomial distribution $\mathsf{PB}(n, p_J)$, where $n \in \Omega(\kappa)$ and for each p_j, it holds that $p_j \in [1/2 - \theta, 1/2 + \theta]$ where $\theta \in (0, 1/2)$ is a constant independent of κ. Then, for any constant ϵ and $s = \Theta(\lg\lg\kappa)$, there are $a, b \in [n]$ such that*

- *For any* $\ell \in [a,b]$, $e^{-\epsilon} \leq \frac{\Pr[\mathsf{PB}(n,p_J)=\ell]}{\Pr[\mathsf{PB}(n,p_J)+s=\ell]} \leq e^{\epsilon}$.
- *For any* $\ell \notin [a,b]$, $\Pr[\mathsf{PB}(n,p_J) = \ell] = \mathsf{negl}(\kappa)$ *and* $\Pr[\mathsf{PB}(n,p_J) + s = \ell] = \mathsf{negl}(\kappa)$.

The proof is found in Appendix D.

8 Highlights of Proof of Theorem 4

Here, we highlight only the important parts of the proof of Theorem 4. The full proof can be found in Appendix E. We show that $\mathcal{F}_{\mathsf{minH}}$ satisfies DDP even when the size of the universe $\mathcal{U}$ of size $n_R \cdot \ell$ is polynomial in κ with $\ell = \Omega(n_R^3)$, and the secret set R is chosen from the uniform distribution on $\mathcal{U}$, conditioned on arbitrary leakage on R of length L, where $L \leq n_R(\lg \ell - 3\lg n_R - 2)$. WLOG, we assume that the adversary corrupts P_1.

We set $n'_R := n_R/3$; looking forward, it is the size of a subset $R' \subset R$, each of whose elements has high remaining min-entropy even after leakage (that we will define in the proof) is considered.

8.1 Min-hash Graph

Consider running the min-hash protocol $\mathcal{F}_{\mathsf{minH}}$ with k iterations such that k_g of them belong to G_θ. For this, we consider all the hash outputs in two different stages and define the following sets:

$$H_1 = \{h_j(A_{+x^*})\}_{j=1}^k, \; H_2 = \{h_j(\mathcal{U} \setminus A_{+x^*})\}_{j=1}^k.$$

Since we are in the random oracle model, each hash value is chosen uniformly at random. For our analysis, we construct the following bipartite graph $(\mathcal{X}, \mathcal{Y}, \mathcal{E})$, which we call the *min-hash graph*, based on the sets A, I and x^* along with the hash functions as follows:

MinhashG$_{H_1}(A, I, x^*, H_2)$**:**

1. Set $\mathcal{X} = \mathcal{U} \setminus A_{+x^*}$. In other words, the graph considers *all potential elements that could be in* $R = B \setminus I$. A distribution of R is equivalent to a distribution of how to choose n_R nodes from $\mathcal{X}$. Note that H_1 determines G_θ (based on the hash values of A and I). We set $\mathcal{Y} = G_\theta$. In other words, $\mathcal{Y}$ corresponds to all the good iterations that could potentially positively contribute to the final count.

2. Let $p_j = \min h_j(I)$. Use H_2 to determine the set of edges:

$$\mathcal{E} = \{(i,j) : (i,j) \in \mathcal{X} \times \mathcal{Y} \text{ and } h_j(x_i) < p_j = \min h_j(I)\}.$$

   In other words, existence of an edge (i,j) means that if node i belongs to R, iteration j will *not* contribute to the final count.

3. Output the resulting bipartite graph $(\mathcal{X}, \mathcal{Y}, \mathcal{E})$.

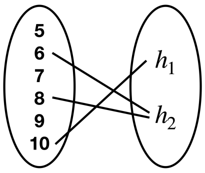

Figure 7: Min-hash graph

	1	2	3	4	5	6	7	8	9	10	11
h_1	0.83	0.25	0.77	0.85	0.93	0.35	0.86	0.92	0.49	0.21	0.5
h_2	0.62	0.83	0.27	0.59	0.63	0.26	0.4	0.26	0.72	0.36	0.6
h_3	0.68	0.11	0.67	0.29	0.82	0.3	0.62	0.23	0.67	0.35	0.7
h_4	0.02	0.43	0.22	0.58	0.69	0.67	0.93	0.56	0.11	0.42	0.8

Table 1: Example Hash Functions

Example. Let the universe be $\mathcal{U} = [11]$. Let $A = \{1,2,3,4\}$, $I = \{2,3\}$, $x^* = 11$. Let the threshold range for the θ-good iterations be $[0.2, 0.7]$. Assume that our protocol runs in 4 iterations using the hash functions defined in table 1.

Figure 7 shows the constructed min-hash graph. We have $\mathcal{X} = \{5, 6, \ldots, 10\}$ and $\mathcal{Y} = \{h_1, h_2\}$; h_3 has been ruled out since $p_3 = h_3(2) = 0.11 \notin [0.2, 0.7]$, and h_4 has been ruled out because $\min h_4(A) \neq \min h_4(I)$. Moreover, we have $p_1 = h_1(2) = 0.25$ and $p_2 = h_2(3) = 0.27$. Note that $(8, h_2) \in \mathcal{E}$, because $h_2(8) < p_2$.

8.2 Fixed Subsets (R', T) of Secret Items and Good Iterations

We fix H_1 and thereby the nodes $\mathcal{X}$ and $\mathcal{Y}$ of the min-hash graph. In this section, as the first step, we fix subsets $R' \subset \mathcal{X}$ and $T \subseteq \mathcal{Y}$ and analyze the noise over the choice of H_2. In other words, we are treating H_2 as private the adversary. Extending this, in the next section, we will consider the actual protocol setting where the hash functions are public and then analyze the noise over *a distribution of* R'.

Edge distribution in the min-hash graph. The probability (over the choice of H_2) that an edge (i, j) forms is exactly equal to p_j. Moreover, since we are in the random oracle model, the probability that (i, j) forms is independent of the probability that any other edge in the graph forms.

Noise distribution. We are interested in *the probability $E^{R'}_{T,r}$ over the choice of H_2 that the final count is reduced by exactly r due to the elements of R' over a bundle T of iterations.* In the random oracle model, the probability depends only on the size of the sets $n' = |R'|$ and $k_b = |T|$. Therefore, we will often use the notation $E^{n'}_{k_b,r} = E^{R'}_{T,r}$. We will sometimes even omit k_b and write $E^{n'}_{r}$. Observe that $E^{n'}_{r}$ is another way of representing an Additive Poisson Binomial distribution. That is, $E^{n'}_{k_b,r} = \Pr[\mathsf{PB}(k_b, p_J) = r]$. Therefore, based on Lemma 6, we have the following:

Corollary 1. *Let $k_b \in \Omega(\kappa)$, and consider any H_1 that makes $|\mathcal{Y}| > k_b$ in the min-hash graph construction. For any $s = O(\lg \lg \kappa)$, any constant ϵ, there are $a, b \in [k_b]$ such that over the choice of H_2, we have*

- *For any $r \notin [a + s, b]$, then $E^{n'_R}_{k_b,r}$ and $E^{n'_R}_{k_b,r-s}$ are both negligible in κ.*
- *For any $r \in [a, b]$, then it holds $e^{-\epsilon/3} \leq E^{n'_R}_{k_b,r} / E^{n'_R}_{k_b,r-s} \leq e^{\epsilon/3}$.*

The above indicates that the distribution over r is amenable for use as a noise distribution in a differential privacy context.

8.3 Noise Over the Choice of R' with Public Hash Functions

Our main technical challenge is to show that the properties needed for differential privacy hold *even when the hash functions are public.*

For this, we first fix H_1 and H_2. Then, we consider the derived min-hash graph $G = (\mathcal{X}, \mathcal{Y}, \mathcal{E})$. Let $\mathcal{D}$ be the distribution of R'. For any set T of iterations of size k_b and any integer r, let $I_{R',T,r}$ be the indicator random variable that is set to 1 if set R' contributes $-r$ to the total count in the min-hash protocol. We define a random variable $D_{T,r}$ that is *the probability that R' contributes to the noise reduction r over iterations in T*:

$$D_{T,r}(\mathcal{D}) := \Pr_{R' \sim \mathcal{D}}[I_{R',T,r}] = \sum_{R'} \Pr_{R' \sim \mathcal{D}}[R'] \cdot I_{R',T,r}.$$

Conditions for the hash functions. Ideally, we would like to show the following:

> For any fixed H_1 and H_2 and over distribution $\mathcal{D}$, it holds that $D_{T,r}$ and $D_{T,r-1}$ (and ultimately $D_{T,r-s}$) are close, except with the tail case of r whose probability weight is negligible.

The universal quantifier for H_1 and H_2 in the above can be *slightly relaxed so that the condition holds with all but small probability over the choice of the hash functions*, which can be captured by showing that $D_{T,r}$ is close to its mean $E^{n'_R}_{\hat{k},r}$ (and then applying Corollary 1).

Geometric collision property. This is essentially to show that $D_{T,r}$ is strongly concentrated around its mean. We could try to apply Chernoff bound to show the concentration property, but we cannot do so because $I_{R'_i,T,r}$ and $I_{R'_j,T,r}$ are not necessarily independent if $R'_i \cap R'_j \neq \emptyset$. Therefore, we instead use Chebyshev for bounding the tail, which requires $D_{T,r}$ to have small variance. Thus, our next goal is to upperbound $\mathsf{Var}[D_{T,r}]$. To do so, we introduce a property of distributions $\mathcal{D}$ over sets R' which we call the "Geometric Collision Property". In a nutshell, this property states that the probability that two sets R'_1, R'_2 drawn independently from $\mathcal{D}$ have intersection of size z is at most $(\frac{1}{n^{0.5}})^z$ for all $z \in [n']$. We show that $\mathsf{Var}[D_{T,r}]$ can be bounded for any distribution over sets R' that has this property.

Definition 9 (Geometric Collision Property)**.** Let $\mathcal{D}$ be a distribution over sets R' of size n'_R. We say that $\mathcal{D}$ has the *Geometric Collision Property* if for all $z \in [n'_R]$

$$\Pr_{R'_i, R'_j \sim \mathcal{D}} \left[|R'_i \cap R'_j| = z \right] \leq \left(\frac{1}{\sqrt{n_R}} \right)^z .$$

Based on this property, we can show the following lemma.

Lemma 7. *Let $k_b \in \Omega(\kappa)$, and consider any H_1 that makes $|\mathcal{Y}| > k_b$ in the min-hash graph construction. Let $\mathcal{D}$ be a distribution over sets of size n'_R with geometric collision property. For any set T of size $k_b \in \Omega(\kappa)$, there exist $a, b \in [k_b]$, such that with probability $1 - O(\frac{k_b \cdot \lg^3(\kappa)}{\sqrt{n_R}})$ over choice of H_2, the following holds:*

- *For all $r \notin [a+s, b]$, $D_{T,r}$ is negligible, where $s = O(\lg\lg\kappa)$.*
- *For all $r \in [a, b]$, $e^{-\epsilon/3} E^{n'_R}_{k_b,r} \leq D_{T,r} \leq e^{\epsilon/3} E^{n'_R}_{k_b,r}$.*

The proof is found in Appendix F.

Multiple bundles of iterations towards DDP with negligible δ. We are not quite done yet. Using the above lemma, we are only able to reduce the failure probability only to $\sim 1/\sqrt{n}$, whereas we would like the failure probability to be negligible. In order to do that, we split the "good" iterations into u bundles, where u is a small superconstant number $u = \lg\lg\kappa$, and argue that with overwhelming probability at least one bundle serves as a good noise. Note that hash outputs are independent in each bundle and so the probability that all u bundles fail should be $(\frac{1}{\sqrt{n}})^u$, which is negligible. For this, we set the parameter $k_b = k_g/u$, where k_g is the number of good iterations.

8.4 Geometric Collision Property In the Face of Leakage

We conclude the proof by showing that R' indeed has the geometric collision property. It is not hard to see that the uniform distribution over all sets R' of size n' from a universe of size $n' \cdot \ell$ (where $\ell \in \Omega(n^3)$) satisfies the "Geometric Collision Property". It would seem, therefore, that we could take this as our secret distribution and the analysis would be complete. Unfortunately, even for the case in which the distribution is sets of size n' chosen uniformly at random from the universe, the analysis is not straightforward. The difficulty stems from the fact that the "noise" in the protocol is tied to the input itself. Therefore, if information about the input is leaked in any other part of the protocol, then the noise distribution changes and may no longer satisfy the required properties. Specifically in our case, learning the number of matches across the two parties' sets with respect to some of the hash functions leaks information about the secret set of the honest party (since the secret set affects those counts).

Strong chain-rule for min-entropy. We first observe that our initial min-entropy in the distribution over secret sets $\mathcal{R}$ is high (approximately $\frac{8n}{9}\lg\ell + 2n$) and that the entire information leaked about R from the counts of the iterations that are not θ-good is small. We can lower-bound the remaining min-entropy in $\mathcal{D}$, therefore, using the weak chain rule for min-entropy [DORS08, Lemma 2.2].

If we want to use the weak chain rule to lower bound the remaining min-entropy with all but $2^{-\kappa}$ probability, however, we need to take a hit of κ in the min-entropy. Recall that each individual element in R can be viewed as being chosen from a set of size ℓ and thus has min-entropy of at most $\lg(\ell) \ll \kappa$. Thus, after applying the weak chain rule and losing more than κ bits of min-entropy, we can have certain elements that have only constant min-entropy, thus implying that collisions are likely in those positions. So the weak chain rule, while leaking only a small number of bits overall, can ruin the geometric collision property. Even worse, the min-entropy definition doesn't rule out the case in which *all* elements of R (i.e. the marginal distributions over each element in R) have only constant min-entropy, while the total min-entropy in R remains high!

This phenomenon has been previously observed and studied in the literature [Skó19]. One way to deal with such a counter-intuitive situation is to actually leak a small amount of *additional* information, known as "spoiled" bits. This will lower the total min-entropy in R, but will ensure that a large fraction of blocks in R still have high min entropy of at least $1.5\lg(n)$. We extend the techniques of [Skó19] to produce spoiling leakage so that the min-entropy in R still stays high in our protocol. We discuss more details about the strong chain rule in the next section.

8.5 Strong Chain Rule

Our strong chain rule considers min-entropy where leakage functions $\ell_1(\cdot), \dots, \ell_n(\cdot)$ are additionally considered. We describe our theorem in a general way, and we hope that it may find future applications in leakage-resilient cryptography.

Sequence of random variables. Recall that R is the set of secret items in the min-hash protocol. Here, we treat R as a sequence of block-by-block random variables $R = (R_1, \dots, R_n)$, associated with (potentially randomized) leakage functions $\ell_1(\cdot), \dots, \ell_n(\cdot)$ with randomness $\rho_1, \dots, \rho_n$. You can think of the blocks as coming in a streaming fashion in order of $R_1, R_2, \dots, R_n$.

Leakage functions. Loosely speaking, the properties we require of the leakage functions are that the i-th leakage ℓ_i can be computed given (R_i, ρ_i), and all the outputs of $(\ell_{i+1}, \ell_{i+2}, \dots, \ell_n)$. We also require that the total number of valid sequences of leakages from $\ell_1(\cdot), \dots, \ell_n(\cdot)$ should be sufficiently small (see Property 1 in Theorem 5 below).

Spoiling functions. Our theorem below states the existence of a spoiling function $f(\cdot)$ with certain properties, as well as properties of the random variables $(R_1, \dots, R_n)$ and $(\rho_1, \dots, \rho_n)$ conditioned on the output of the spoiling function $f(R)$.

The properties of $(R_1, \dots, R_n)$ and $(\rho_1, \dots, \rho_n)$ are roughly the following: (1) There exist disjoint sets V, W such that $V \cup W = [n]$ that are determined by $f(R)$. (2) Blocks $\{R_i\}_{i \in V}$ have high min-entropy conditioned on $f(R)$. (3) Blocks $\{R_i\}_{i \in W}$ have small support size (low max-entropy) conditioned on $f(R)$. (4) For $i \in V$, the random strings ρ_i are uniform random and independent conditioned on $f(R)$. (See Properties (5)-(8) in Theorem 5).

The properties of $f(\cdot)$ are roughly the following: (1) The failure probability (outputting $\perp$) is small. (2) As long as the total number of valid sequences of leakages from $\ell_1(\cdot), \dots, \ell_n(\cdot)$ is sufficiently small, the image size of f is small. This property ensures that we do not lose too much of the total min-entropy of R by releasing $f(R)$ (3) The leakages $\{\ell_i(\cdot)\}_{i\in W}$ can be computed given $f(R)$. (See Properties (2)-(4) in Theorem 5)

Our theorem. The main difference between our spoiling lemma and prior ones is that our min and max entropy guarantees on $R = (R_1, \ldots, R_n) \mid f(R)$ hold even with respect to additional leakage $\{\ell_i\}_{i\in W}$ which is included in the spoiled bits $f(R)$.

Theorem 5 (Block structures with few bits spoiled and leakage)**.** *Let $\mathcal{U} = U_1 \times \cdots \times U_n$ be a fixed universe and $R = (R_1, \ldots, R_n)$ be a sequence of (possibly correlated) random variables where each R_i is over U_i (and all are disjoint) and $|U_i| = \ell$ for all i. Let $\rho_1, \ldots, \rho_n$ be a sequence of uniformly random strings over $\{0,1\}^m$ and let $\ell_1(\cdot), \ldots, \ell_n(\cdot)$ be leakage functions. Then, for any $\epsilon \in (0,1)$, any $\delta > 0$ and any $c \in [2^\delta, \ell/2^\delta]$, there exists a spoiling leakage function $f(R)$ that satisfies the following properties.*

1. *A sequence $\beta_1, \ldots, \beta_n$ is valid if for all $i \in V$, $\beta_i = \perp$ and for all $i \in W$, $\beta_i = \ell_i(R_i, \rho_i, \beta_{>i})$, where $\beta_{>i} = (\beta_{i+1}, \ldots, \beta_n)$. We require that the number of valid sequences $\beta_1, \ldots, \beta_n$ is at most B.*
2. *It holds that $\Pr_R[f(R) = \perp] \le \epsilon n$.*
3. $|Im(f)| \le B \cdot (2(\lg(\ell) + \lg(1/\epsilon))/\delta)^n$.
4. *Conditioned on any $y \in Im(f) \setminus \{\perp\}$, for all $i \in W$, the leakage $\ell_i(R_i, \rho_i, \beta_{>i})$ can be computed from y. Here, $\beta_j = \perp$ if $j \in V$ and $\beta_j = \ell_j(R_j, \rho_j, \beta_{>j})$ otherwise.*
5. *Let $Im(f)$ be the set of images of f. Every $y \in Im(f) \setminus \{\perp\}$ specifies two disjoint sets V and W such that $V \cup W = [n]$.*
6. *Conditioned on any $y \in Im(f) \setminus \{\perp\}$, for every $i \in V$, every element in distribution $R_i \mid R_{<i}$ has low probability weight, i.e.,*
$$\forall y \in Im(f) \setminus \{\perp\}, \forall r \text{ s.t. } f(r) = y, \forall i \in V: \quad \Pr\left[R_i = r_i \;\middle|\; R_{<i} = r_{<i},\ y\right] \le \frac{2^\delta}{c}.$$
7. *Conditioned on any $y \in Im(f) \setminus \{\perp\}$, for every $i \in W$, it holds that $R_i \mid R_{<i}$ has small support size, i.e.,*
$$\begin{gathered}\forall y \in Im(f) \setminus \{\perp\}, \forall r \text{ s.t. } f(r) = y, \forall i \in W: \\ |\{r_i : \Pr[R_i = r_i | R_{<i} = r_{<i}, y]] \ge 0\}| \le 2^\delta \cdot c.\end{gathered}$$
8. *$\{\rho_i\}_{i\in V}$ are distributed independently and uniformly at random conditioned on $f(R)$.*

The proof is found in Appendix G. Typically, one would like to set c as large as possible, while ensuring that the size of V remains above some threshold. The achievable tradeoffs between c and $|V|$ are determined by the min-entropy of R before the spoiling bits $f(R)$ are released. For our applications, we require $c = n^{1.5}$ and $|V| \ge n/3$. In Section H, We show that our min-entropy assumption on R implies that this parameter setting is achievable.

9 Empirical Evaluation

9.1 Comparison With Prior Work

We compare our noisy min-hash (NMH) protocol π_{NMH} with the current state-of-the-art approach, called sketch-flip-merge (SFM) [HTC23] and the generalized randomized response mechanism (GRR) [ABS20]. In particular, we evaluate the trade-off between communication cost and cardinality estimation accuracy, while achieving (almost) the same level of privacy guarantee as follows:

- For a given privacy parameter ϵ (with δ fixed to 2^{-40}), we choose the right amount of noise for our protocol and vary the number of hash functions k to measure the communication cost and estimation accuracy trade-off.
- We then compare these accuracy results with the state-of-the-art protocols using the same communication and privacy parameter.

  We use the relative root mean squared error (RRMSE) of the union size as our accuracy metric. This choice is primarily to ensure a fair comparison between our protocol and the SFM protocol. Further details on this matter are provided in the discussion of the SFM protocol below.

  In Figure 9, we demonstrate the comparison of NMH, SFM, and GRR.

Our protocol. We calculate the communication of our protocol π_{NMH} with $\mathcal{F}_{\mathsf{psi\text{-}ca}}$ instantiated with the PSI-CA protocol described in Figure 8. It is a variant of the protocol in [CGT12], where H_2 is applied to $\{a'_i\}_{i\in[v]}$ and $\{b'_j\}_{j\in[w]}$ in order to reduce the communication. The original protocol is secure under the DDH assumption in the random oracle model. Essentially the same security proof found in the original paper can be applied to show the security of this variant, when H_2 is also modeled as a random oracle.

Private Set Intersection Cardinality

Let G be a multiplicative group of order q. Let $H_1 : \{0,1\}^* \to G$ and $H_2 : \{0,1\}^* \to \{0,1\}^\lambda$ be hash functions.
Input: P_1 has $C = \{c_1, \ldots, c_v\}$ and P_2 has $S = \{s_1, \ldots, s_w\}$.

1. P_1 samples a random exponent $R_c \leftarrow \mathbb{Z}_q$. For $i \in [v]$, P_1 computes $a_i = H_1(c_i)^{R_c}$. P_1 sends $(a_1, \ldots, a_v)$.
2. P_2 samples $R_s \leftarrow \mathbb{Z}_q$ and computes $(a'_1, a'_2, \ldots, a'_v) = \mathsf{shuffle}(a_1^{R_s}, \ldots, a_v^{R_s})$. P_2 also computes $(b_1, b_2, \ldots, b_w) = \mathsf{shuffle}(H_1(s_1)^{R_s}, \ldots, H_1(s_w)^{R_s})$. P_2 sends $(H_2(a'_1), \ldots, H_2(a'_w))$ and $(b_1, \ldots, b_w)$ to P_1.
3. P_1 computes $(b'_1, \ldots, b'_w) = (b_1^{R_c}, \ldots, b_w^{R_c})$. P_1 outputs the following value:

$$|\ \{H_2(a'_1), \ldots, H_2(a'_v)\} \cap \{H_2(b'_1), \ldots, H_2(b'_w)\}|.$$

Figure 8: PSI-CA Protocol.

We briefly sketch the security proof here while referring the full proof to the original paper [CGT12]. We first show the simulator for the corrupted P_1. Let t be the protocol output (i.e., set intersection cardinality). The simulator chooses random t indices $(i_1, \ldots, i_t)$ (resp., $(j_1, \ldots, j_t)$) from $[v]$ (resp., $[w]$). In order to prepare $(b_1, \ldots, b_w)$, the simulator replaces $H_1(s_{j_k})^{R_s}$ (for $k \in [t]$) with $H_1(c_{i_k})^{R_s}$, and the remaining values $H_1(s_h)^{R_s}$ are simulated with random numbers. Since H_1 is a random oracle (i.e., for an input x, we have $H_1(x) = g^r$ for a random r), this simulation is indistinguishable under the DDH assumption. When P_2 is corrupted, the first message $\{a_i = H_1(c_i)^{R_c} : i \in [v]\}$ is simulated by random values. The simulation is also indistinguishable under the DDH assumption. The above PSI-CA protocol exchanges $v + w$ elliptic curve points and w hashes, resulting in $(v + w) \cdot 256 + w \cdot 80$ bits. In protocol π_{NMH}, the parties will run this PSI-CA protocol by setting $v = w = k + 2\ell_B$.

Sketch-Flip-Merge (SFM) [HTC23]. While our main focus is on comparing the accuracy of Jaccard Index estimation, in the absence of available code, we had to rely on their analysis of the relative root mean squared error (RRMSE) of cardinality estimation instead of Jaccard Index estimation. This poses challenges in evaluating the accuracy of the Jaccard Index for SFM. In particular, although the Jaccard Index can be estimated by calculating the ratio of estimated intersection size over the estimated union size, its RRMSE cannot be directly calculated from RRMSEs for the intersection and union sizes. This is because the two estimates have dependency, and we can only conjecture that the derived estimate through the division operation will probably have a worse RRMSE.

In the end, giving a slight advantage to SFM, we decided to focus on the accuracy of cardinality estimation of the size of the union only. In our case, the union size was estimated based on the Jaccard Index from the min-hash protocol and n_A and n_B. Following the approach of SFM [HTC23], we perform $m = 1000$ estimates to measure the accuracy in the form of relative root mean squared error (RRMSE); that is, letting $\hat{n}_{U,1}, \ldots, \hat{n}_{U,m}$ be the union size estimates, and n_U be the real union size, we define $\mathrm{RRMSE}(\hat{n}_{U,1}, \ldots, \hat{n}_{U,m}; n_U)$ to be $\frac{1}{n_U}\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{n}_{U,i} - n_U)^2}$. To match the communication complexity, we set the sketch of SFM to be a $(B \times P)$-bit matrix such that $B \cdot P = 592w$ and $P = 24$.

Generalized Randomized Response (GRR) [ABS20]. We also compare our protocol with the generalized randomized response MinHash protocol in [ABS20]. Following their guidance in experiments, we select the range of their hash function to be a single bit and let their protocol use $592w$ hash functions to match our communication cost.

Since their actual protocol would take too long to run for large n and k, in order to facilitate the large number of hash functions, we wrote code simulating the error based on their privacy and utility analysis. As with the other protocols, we perform 1000 estimates to lower the variance of the errors. To align with our other comparison with SFM, we report the relative root mean squared error of the union size.

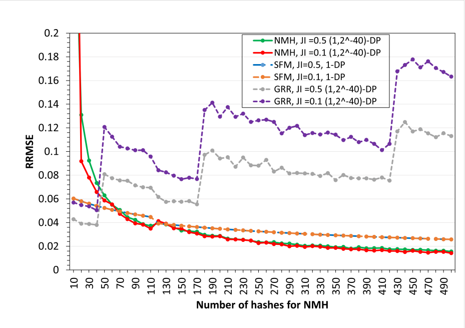

Figure 9: Accuracy: NMH, SFM, GRR.

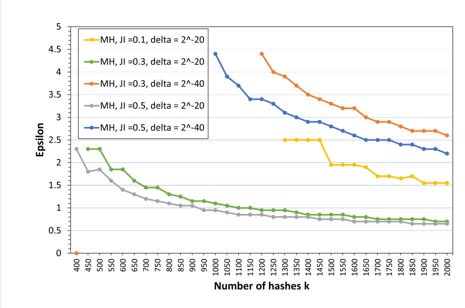

Figure 10: DDP based on k, and JI

Comparison Results. In Figure 9, we demonstrate the comparison of NMH, SFM, and GRR. We set $n = 10^6$. The result shows that our error is consistently smaller than both SFM and GRR for a reasonable range of communication costs, which corresponds to the usage of $k \in [100, 500]$ hashes for our noisy min-hash protocol. Specifically, as we increase the number of hash functions, both our protocol and SFM achieve increased accuracy, due to larger sketch sizes that better represent the input sets. On the other hand, while GRR performs well with smaller communication, adding more communication becomes counter productive. This is because each additional bit in their protocol corresponds to an extra hash function output, which increases the noise needed to achieve the same privacy guarantee. Finally, both GRR and our protocol exhibit spikes in the accuracy trends, corresponding to crossover points where a significant amount of additional noise is required to keep δ from getting above 2^{-40}.

In concluding remarks, it is noteworthy that both SFM and GRR protocol discloses the entire noisy sketch, revealing collective information about the party's input set. We note that differential privacy does not prohibit revealing collective information about the inputs; rather, it mandates that individual contributions should not be discernible in the output. In contrast, our min-hash protocol employs secure two-party computation and discloses no information about the input set, except for the final $(\lg k)$-bit output. When deciding which scheme to use, depending on the specific use case, this observation may need to be taken into account.

9.2 DDP of Noiseless Protocol

We empirically evaluate the DDP guarantee for the noiseless min-hash protocol in the public hash setting, in which each element in the secret set has high entropy. As before we set $n_A = n_B = 10^6$. Figure 10 shows how the privacy parameter ϵ changes with the number k of iterations. Although we demonstrate the results for $JI \leq 0.5$, the results for $JI > 0.5$ are similar. We omit the data points where ϵ is greater than 5, which happens when k is small, and focus on the more meaningful ϵ range. We observe the following:

- Roughly speaking, when the number k of iterations of the min-hash protocol is reasonably large (at least 500), the noiseless min-hash protocol provides a decent level of DDP with privacy parameter $\epsilon \in [0.5, 5]$.
- Higher values of k correspond to improved privacy parameters. Note that as k grows, more iterations will be θ-good. Since hashes of non-intersecting items work as noise in θ-good iterations, more θ-good iterations will essentially amount to adding more noise, therefore offering better privacy guarantee.
- When k is the same, the best privacy parameter is achieved when JI is around 0.5. This is due to the likelihood that a hash function is θ-good being maximized when striking a balance between the two conditions stipulated in Definition 8: (i) the hash of an intersecting item should be the minimum hash value, and (ii) the minimum hash value is neither too large nor too small.

References

[ABS20] Martin Aumüller, Anders Bourgeat, and Jana Schmurr. Differentially private sketches for jaccard similarity estimation. In *SISAP 2020*, LNCS, pages 18–32. Springer, 2020. doi:10.1007/978-3-030-60936-8_2.

[ACSS23] Idan Attias, Edith Cohen, Moshe Shechner, and Uri Stemmer. A framework for adversarial streaming via differential privacy and difference estimators. In Yael Tauman Kalai, editor, *ITCS 2023: 14th Innovations in Theoretical Computer Science Conference*, volume 251, pages 8:1–8:19, Cambridge, MA, USA, January 10–13, 2023. Leibniz International Proceedings in Informatics (LIPIcs). doi:10.4230/LIPIcs.ITCS.2023.8.

[BBDS12] Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The johnson-lindenstrauss transform itself preserves differential privacy. In *53rd Annual Symposium on Foundations of Computer Science*, pages 410–419, New Brunswick, NJ, USA, October 20–23, 2012. IEEE Computer Society Press. doi:10.1109/FOCS.2012.67.

[BCG14] Carlo Blundo, Emiliano De Cristofaro, and Paolo Gasti. Espresso: Efficient privacy-preserving evaluation of sample set similarity. *J. Comput. Secur.*, 22(3):355–381, 2014. doi:10.3233/jcs-130482.

[BGKS13] Raef Bassily, Adam Groce, Jonathan Katz, and Adam Smith. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *54th Annual Symposium on Foundations of Computer Science*, pages 439–448, Berkeley, CA, USA, October 26–29, 2013. IEEE Computer Society Press. doi:10.1109/FOCS.2013.54.

[BGMZ97] Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. Syntactic clustering of the web. *Comput. Networks*, 29(8-13):1157–1166, 1997. doi:10.1016/s0169-7552(97)00031-7.

[BJWY22] Omri Ben-Eliezer, Rajesh Jayaram, David P. Woodruff, and Eylon Yogev. A framework for adversarially robust streaming algorithms. *J. ACM*, 69(2):17:1–17:33, 2022. doi:10.1145/3498334.

[BLV19] Elette Boyle, Rio LaVigne, and Vinod Vaikuntanathan. Adversarially robust property-preserving hash functions. In Avrim Blum, editor, *ITCS 2019: 10th Innovations in Theoretical Computer Science Conference*, volume 124, pages 16:1–16:20, San Diego, CA, USA, January 10–12, 2019. Leibniz International Proceedings in Informatics (LIPIcs). doi:10.4230/LIPIcs.ITCS.2019.16.

[BNO08] Amos Beimel, Kobbi Nissim, and Eran Omri. Distributed private data analysis: Simultaneously solving how and what. In *Crypto 2008*, pages 451–468. Springer, 2008. doi:10.1007/978-3-540-85174-5_25.

[BNSGT17] Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Guha Thakurta. Practical locally private heavy hitters. *Advances in Neural Information Processing Systems*, 30, 2017. doi:10.5555/3294771.3294989.

[BO13] Gilles Barthe and Federico Olmedo. Beyond differential privacy: Composition theorems and relational logic for f-divergences between probabilistic programs. In Fedor V. Fomin, Rusins Freivalds, Marta Z. Kwiatkowska, and David Peleg, editors, *ICALP 2013: 40th International Colloquium on Automata, Languages and Programming, Part II*, volume 7966 of *Lecture Notes in Computer Science*, pages 49–60, Riga, Latvia, July 8–12, 2013. Springer Berlin Heidelberg, Germany. doi:10.1007/978-3-642-39212-2_8.

[Bro97] A. Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences, International Conference on*, page 21. IEEE Computer Society, 1997. doi:10.1109/sequen.1997.666900.

[BS15] Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 127–135, 2015. doi:10.1145/2746539.2746632.

[CCMS19] T.-H. Hubert Chan, Kai-Min Chung, Bruce M. Maggs, and Elaine Shi. Foundations of differentially oblivious algorithms. In Timothy M. Chan, editor, *30th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2448–2467, San Diego, CA, USA, January 6–9, 2019. ACM-SIAM. doi:10.1137/1.9781611975482.150.

[CDSKY20] Seung Geol Choi, Dana Dachman-Soled, Mukul Kulkarni, and Arkady Yerukhimovich. Differentially-private multi-party sketching for large-scale statistics. *Proceedings on Privacy Enhancing Technologies*, 2020(3):153–174, 2020. doi:10.2478/popets-2020-0047.

[CFGT12] Emiliano De Cristofaro, Sky Faber, Paolo Gasti, and Gene Tsudik. Genodroid: are privacy-preserving genomic tests ready for prime time? In *WPES 2012*, pages 97–108. ACM, 2012. doi:10.1145/2381966.2381980.

[CGT12] Emiliano De Cristofaro, Paolo Gasti, and Gene Tsudik. Fast and private computation of cardinality of set intersection and union. In *CANS 2012*, volume 7712, pages 218–231. Springer, 2012. doi:10.1007/978-3-642-35404-5_17.

[COK22] Wei-Ning Chen, Ayfer Ozgur, and Peter Kairouz. The poisson binomial mechanism for unbiased federated learning with secure aggregation. In *ICML 2022*, volume 162, pages 3490–3506. PMLR, 2022.

[DKZ18] Stefan Dziembowski, Tomasz Kazana, and Maciej Zdanowicz. Quasi chain rule for min-entropy. *Inf. Process. Lett.*, 134:62–66, 2018. doi:10.1016/j.ipl.2018.02.007.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography (TCC 2006)*, pages 265–284. Springer, 2006. doi:10.1007/11681878_14.

[Doe18] Benjamin Doerr. Probabilistic tools for the analysis of randomized optimization heuristics. *CoRR*, abs/1801.06733, 2018. arXiv:1801.06733.

[DORS08] Yevgeniy Dodis, Rafail Ostrovsky, Leonid Reyzin, and Adam Smith. Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. *SIAM Journal on Computing*, 38(1):97–139, jan 2008. doi:10.1137/060651380.

[DR+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. doi:10.1561/9781601988195.

[DTT22] Charlie Dickens, Justin Thaler, and Daniel Ting. Order-invariant cardinality estimators are differentially private. *Advances in Neural Information Processing Systems*, 35:15204–15216, 2022. doi:10.5555/3600270.3601376.

[Dwo06] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer, 2006. doi:10.1007/11787006_1.

[Fab16] Sky Faber. Variants of privacy preserving set intersection and their practical applications. *PhD Thesis*, 2016.

[FIM+01] Joan Feigenbaum, Yuval Ishai, Tal Malkin, Kobbi Nissim, Martin Strauss, and Rebecca N. Wright. Secure multiparty computation of approximations. In Fernando Orejas, Paul G. Spirakis, and Jan van Leeuwen, editors, *ICALP 2001: 28th International Colloquium on Automata, Languages and Programming*, volume 2076 of *Lecture Notes in Computer Science*, pages 927–938, Heraklion, Crete, Greece, July 8–12, 2001. Springer Berlin Heidelberg, Germany. doi:10.1007/3-540-48224-5_75.

[FLS22] Nils Fleischhacker, Kasper Green Larsen, and Mark Simkin. Property-preserving hash functions for hamming distance from standard assumptions. In Orr Dunkelman and Stefan Dziembowski, editors, *Advances in Cryptology – EUROCRYPT 2022, Part II*, volume 13276 of *Lecture Notes in Computer Science*, pages 764–781, Trondheim, Norway, May 30 – June 3, 2022. Springer, Cham, Switzerland. doi:10.1007/978-3-031-07085-3_26.

[FS21] Nils Fleischhacker and Mark Simkin. Robust property-preserving hash functions for hamming distance and more. In Anne Canteaut and François-Xavier Standaert, editors, *Advances in Cryptology – EUROCRYPT 2021, Part III*, volume 12698 of *Lecture Notes in Computer Science*, pages 311–337, Zagreb, Croatia, October 17–21, 2021. Springer, Cham, Switzerland. doi:10.1007/978-3-030-77883-5_11.

[GKLX22] S. Dov Gordon, Jonathan Katz, Mingyu Liang, and Jiayu Xu. Spreading the privacy blanket: - differentially oblivious shuffling for differential privacy. In Giuseppe Ateniese and Daniele Venturi, editors, *ACNS 22: 20th International Conference on Applied Cryptography and Network Security*, volume 13269 of *Lecture Notes in Computer Science*, pages 501–520, Rome, Italy, June 20–23, 2022. Springer, Cham, Switzerland. doi:10.1007/978-3-031-09234-3_25.

[GRR19] Adam Groce, Peter Rindal, and Mike Rosulek. Cheaper private set intersection via differentially private leakage. *Proc. Privacy Enhancing Technologies (PETS)*, 2019(3):6–25, 2019. doi:10.2478/popets-2019-0034.

[HKKN01] Shai Halevi, Robert Krauthgamer, Eyal Kushilevitz, and Kobbi Nissim. Private approximation of NP-hard functions. In *33rd Annual ACM Symposium on Theory of Computing*, pages 550–559, Crete, Greece, July 6–8, 2001. ACM Press. doi:10.1145/380752.380850.

[HLTW22] Justin Holmgren, Minghao Liu, LaKyah Tyner, and Daniel Wichs. Nearly optimal property preserving hashing. In Yevgeniy Dodis and Thomas Shrimpton, editors, *Advances in Cryptology – CRYPTO 2022, Part III*, volume 13509 of *Lecture Notes in Computer Science*, pages 473–502, Santa Barbara, CA, USA, August 15–18, 2022. Springer, Cham, Switzerland. doi:10.1007/978-3-031-15982-4_16.

[HMFS17] Xi He, Ashwin Machanavajjhala, Cheryl J. Flynn, and Divesh Srivastava. Composing differential privacy and secure computation: A case study on scaling private record linkage. In Bhavani M. Thuraisingham, David Evans, Tal Malkin, and Dongyan Xu, editors, *ACM CCS 2017: 24th Conference on Computer and Communications Security*, pages 1389–1406, Dallas, TX, USA, October 31 – November 2, 2017. ACM Press. doi:10.1145/3133956.3134030.

[HQYC22] Ziyue Huang, Yuan Qiu, Ke Yi, and Graham Cormode. Frequency estimation under multiparty differential privacy: one-shot and streaming. *Proc. VLDB Endow.*, 15(10):2058–2070, June 2022. doi:10.14778/3547305.3547312.

[HTC23] Jonathan Hehir, Daniel Ting, and Graham Cormode. Sketch-flip-merge: Mergeable sketches for private distinct counting. In *ICML 2023*, volume 202, pages 12846–12865. PMLR, 2023. doi:10.5555/3618408.3618930.

[Jac01] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901. doi:10.5169/seals-266450.

[JKWC22] Bo Jiang, Hamid Krim, Tianfu Wu, and Derya Cansever. Refining self-supervised learning in imaging: Beyond linear metric. In *ICIP 2022*, pages 76–80. IEEE, 2022. doi:10.1109/icip46576.2022.9897745.

[KWS+20] Benjamin Kreuter, Craig William Wright, Evgeny Sergeevich Skvortsov, Raimundo Mirisola, and Yao Wang. Privacy-preserving secure cardinality and frequency estimation. Technical report, Google, LLC, 2020.

[LLSS19] Tian Li, Zaoxing Liu, Vyas Sekar, and Virginia Smith. Privacy for free: Communication-efficient learning with differential privacy using sketches. *arXiv preprint arXiv:1911.00972*, 2019. doi:10.48550/arXiv.1911.00972.

[LOZ12] Ping Li, Art B. Owen, and Cun-Hui Zhang. One permutation hashing. In *In Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012.*, pages 3122–3130, 2012. doi:10.5555/2999325.2999482.

[MDDC16] Luca Melis, George Danezis, and Emiliano De Cristofaro. Efficient private statistics with succinct sketches. In *Proceedings 2016 Network and Distributed System Security Symposium*, NDSS 2016. Internet Society, 2016. URL: http://dx.doi.org/10.14722/ndss.2016.23175, doi:10.14722/ndss.2016.23175.

[MG18] Sahar Mazloom and S. Dov Gordon. Secure computation with differentially private access patterns. In David Lie, Mohammad Mannan, Michael Backes, and XiaoFeng Wang, editors, *ACM CCS 2018: 25th Conference on Computer and Communications Security*, pages 490–507, Toronto, ON, Canada, October 15–19, 2018. ACM Press. doi:10.1145/3243734.3243851.

[MLRG20] Sahar Mazloom, Phi Hung Le, Samuel Ranellucci, and S. Dov Gordon. Secure parallel computation on national scale volumes of data. In Srdjan Capkun and Franziska Roesner, editors, *USENIX Security 2020: 29th USENIX Security Symposium*, pages 2487–2504. USENIX Association, August 12–14, 2020.

[MMNW11] Darakhshan Mir, Shan Muthukrishnan, Aleksandar Nikolov, and Rebecca N Wright. Pan-private algorithms via statistics on sketches. In *ACM SIGMOD-SIGACT-SIGART*, pages 37–48, 2011. doi:10.1145/1989284.1989290.

[NvVT20] Saskia Nuñez von Voigt and Florian Tschorsch. Rrtxfm: Probabilistic counting for differentially private statistics. In *Digital Transformation for a Sustainable Society in the 21st Century: I3E 2019 IFIP WG 6.11 International Workshops*, pages 86–98. Springer, 2020. doi:10.1007/978-3-030-39634-3_9.

[PS21] Rasmus Pagh and Nina Mesing Stausholm. Efficient Differentially Private F_0 Linear Sketching. In Ke Yi and Zhewei Wei, editors, *24th International Conference on Database Theory (ICDT 2021)*, volume 186 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 18:1–18:19, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICDT.2021.18, doi:10.4230/LIPIcs.ICDT.2021.18.

[PT22] Rasmus Pagh and Mikkel Thorup. Improved utility analysis of private countsketch. *Advances in Neural Information Processing Systems*, 35:25631–25643, 2022. doi:10.5555/3600270.3602128.

[RCS$^+$19] M. Sadegh Riazi, Beidi Chen, Anshumali Shrivastava, Dan Wallach, and Farinaz Koushanfar. Sub-linear privacy-preserving near-neighbor search. Cryptology ePrint Archive, Report 2019/1222, 2019. URL: https://eprint.iacr.org/2019/1222.

[SCRS17] Elaine Shi, T.-H. Hubert Chan, Eleanor Gilbert Rieffel, and Dawn Song. Distributed private data analysis: Lower bounds and practical constructions. *ACM Trans. Algorithms*, 13(4):50:1–50:38, 2017. doi:10.1145/3146549.

[Skó19] Maciej Skórski. Strong chain rules for min-entropy under few bits spoiled. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1122–1126. IEEE, 2019. doi:10.1109/isit.2019.8849240.

[SNY17] Rade Stanojevic, Mohamed Nabeel, and Ting Yu. Distributed cardinality estimation of set operations with differential privacy. In *2017 IEEE Symposium on Privacy-Aware Computing (PAC)*, pages 37–48. IEEE, 2017. doi:10.1109/pac.2017.43.

[SSGT20] Adam Smith, Shuang Song, and Abhradeep Guha Thakurta. The flajolet-martin sketch itself preserves differential privacy: Private counting with minimal space. *NeurIPS 2020*, 33:19561–19572, 2020. doi:10.5555/3495724.3497365.

[STS18] Hagen Sparka, Florian Tschorsch, and Björn Scheuermann. P2kmv: A privacy-preserving counting sketch for efficient and accurate set intersection cardinality estimations. *Cryptology ePrint Archive*, 2018. doi:10.14279/DEPOSITONCE-8374.

[SV15] Igal Sason and Sergio Verdú. Bounds among f-divergences. *CoRR*, abs/1508.00335, 2015. arXiv:1508.00335.

[TBK07] Chayant Tantipathananandh, Tanya Y. Berger-Wolf, and David Kempe. A framework for community identification in dynamic social networks. In *KDD*, pages 717–726. ACM, 2007. doi:10.1145/1281192.1281269.

[TL24] Yang Tan and Bo Lv. Break two psi-ca protocols in polynomial time. In *Proceedings of the 2024 16th International Conference on Machine Learning and Computing*, ICMLC '24, page 65–72, New York, NY, USA, 2024. Association for Computing Machinery. doi:10.1145/3651671.3651682.

[WPS22] Lun Wang, Iosif Pinelis, and Dawn Song. Differentially private fractional frequency moments estimation with polylogarithmic space. In *International Conference on Learning Representations*, 2022.

[WWT$^+$19] Huijun Wu, Chen Wang, Yuriy Tyshetskiy, Andrew Docherty, Kai Lu, and Liming Zhu. Adversarial examples for graph data: Deep insights into attack and defense. In *IJCAI 2019*, pages 4816–4823. ijcai.org, 2019. doi:10.24963/ijcai.2019/669.

[YLL$^+$17] Ziqi Yan, Jiqiang Liu, Gang Li, Zhen Han, and Shuo Qiu. Privmin: Differentially private minhash for jaccard similarity computation. *CoRR*, abs/1705.07258, 2017. arXiv:1705.07258, doi:10.48550/arXiv.1705.07258.

[YWR$^+$19] Ziqi Yan, Qiong Wu, Meng Ren, Jiqiang Liu, Shaowu Liu, and Shuo Qiu. Locally private jaccard similarity estimation. *Concurr. Comput. Pract. Exp.*, 31(24), 2019. doi:10.1002/cpe.4889.

[ZQR$^+$22] Fuheng Zhao, Dan Qiao, Rachel Redberg, Divyakant Agrawal, Amr El Abbadi, and Yu-Xiang Wang. Differentially private linear sketches: Efficient implementations and applications. *NeurIPS 2022*, 35:12691–12704, 2022. doi:10.5555/3600270.3601192.

[Ös02] Ferdinand Österreicher. Csiszar's f-divergence-basic properties. *RGMIA Research Report Collection*, 2002.

A Proof of Lemma 3

Let $c = p/(1-p)$. We need to find $a < np$ and $b > np$ satisfying the following condition.

$$\frac{\Pr_{\mathrm{B}(n,p)}[a]}{\Pr_{\mathrm{B}(n,p)+s}[a]} = \frac{\binom{n}{a} p^a (1-p)^{n-a}}{\binom{n}{a-s} p^{a-s} (1-p)^{n-a+s}} < \left(\frac{n-a}{a-s}\right)^s \cdot c^s \leq e^{\epsilon}.$$

$$\frac{\Pr_{\mathrm{B}(n,p)}[b]}{\Pr_{\mathrm{B}(n,p)+s}[b]} = \frac{\binom{n}{b} p^b (1-p)^{n-b}}{\binom{n}{b-s} p^{b-s} (1-p)^{n-b+s}} > \left(\frac{n-b}{b}\right)^s \cdot c^s \geq e^{-\epsilon}.$$

Case 1: $s \geq \epsilon$**.** We set $a = \frac{np+s(1-p)\cdot e^{\epsilon/s}}{e^{\epsilon/s}\cdot(1-p)+p}$ and $b = \frac{e^{\epsilon/s}\cdot np}{(1-p)+e^{\epsilon/s}\cdot p}$. Note these values satisfy the above inequalities.

To show the second requirement of the tail bound, it suffices to show that $\Pr_{\mathrm{B}(n,p)}[X \leq a+s] = \mathsf{negl}(\kappa)$; the other case holds similarly.

Let $\mu := np \in \Theta(\kappa)$ and let $d = 1 - (a+s)/\mu$. By applying the Chernoff bound, we have

$$\Pr_{X \leftarrow \mathrm{B}(n,p)}[X \leq a+s] = \Pr[X \leq (1-d)\mu] \leq \exp(-d^2\mu/2).$$

We will show that we have $d = \Omega(\frac{1}{\lg \lg \kappa})$, which implies that with $\mu \in \Theta(\kappa)$, the above probability is negligible in κ. Let $t = s(1-p)\cdot e^{\epsilon/s}$ and $u = e^{\epsilon/s}\cdot(1-p)+p$. Then, we have $a = \frac{\mu+t}{u}$. Note that we have $\epsilon/s\cdot(1-p)+1 \leq u \leq e$. We have the following:

$$d = \frac{\mu - a - s}{\mu} = \left(1 - \frac{1}{u}\right) - \frac{t/u+s}{\mu} \geq \frac{\epsilon/s\cdot(1-p)}{e} - \frac{t/u+s}{\mu} = \Omega\left(\frac{1}{\lg\lg\kappa}\right) - \tilde{O}(1/\kappa).$$

Case 2: $s \leq \epsilon$**.** Let $a = \frac{c}{c+e}\cdot(n+e\epsilon) < np$ and $b = \frac{c}{c+e^{-1}}\cdot n > np$. Observe that $\frac{n-a}{b-s}\cdot c = e$ and $\frac{n-b}{b}\cdot c = e^{-1}$. Therefore, given $s \leq \epsilon$, the above inequalities hold. The tail bounds specified as the second condition of the lemma can be shown using the Chernoff bound since $np - a \in \Theta(np) = \Theta(\kappa)$ and so does $b - np$.

B Proof of Lemma 4

Let $t = \frac{1}{n_R}$, Let $\mathsf{low} = 1 - \left(\frac{1}{2}+\theta\right)^t$ and $\mathsf{high} = 1 - \left(\frac{1}{2}-\theta\right)^t$. Then, we have:

$$\begin{aligned}
&\Pr_h[\mathsf{good}_\theta(h, A, I, n_B)] \\
&= \Pr_h\left[(\min h(I) \geq \mathsf{low}) \wedge (\min h(I) = \min h(A))\right] \\
&\quad - \Pr_h\left[(\min h(I) \geq \mathsf{high}) \wedge (\min h(I) = \min h(A))\right] \\
&= \Pr_h\left[\min h(A) \geq \mathsf{low}\right] \cdot \Pr\left[\min h(I) = \min h(A) \mid \min h(A) \geq \mathsf{low}\right] \\
&\quad - \Pr_h\left[\min h(A) \geq \mathsf{high}\right] \cdot \Pr\left[\min h(I) = \min h(A) \mid \min h(A) \geq \mathsf{high}\right] \\
&\geq \left(\frac{1}{2}+\theta\right)^{t\cdot n_A} \cdot \frac{n_I}{n_A} - \left(\frac{1}{2}-\theta\right)^{t\cdot n_A} \cdot \frac{n_I}{n_A}
\end{aligned}$$

C Proof of Lemma 5

We can lowerbound $|K_\theta|$ with $\mathrm{B}(k-s, p_\theta)$. Recall $s \in O(\lg\lg\kappa)$. Let $\mu = (k-s)p_\theta = \Omega(\kappa)$. Applying the Chernoff Bound, we have $\Pr\left[|K_\theta| \leq (1-1/3)\mu\right] \leq \exp(-\mu/18) \leq \mathsf{negl}(\kappa)$.

D Proof of Lemma 6

Hockey stick divergence. We first review hockey stick divergence [SV15, Ös02, BO13]. The hockey-stick divergence between two probability measures P, Q over Z is defined as:

$$\mathsf{D}^{\mathsf{hs}}_{\alpha}(X, Y) = \sup_{S \subseteq Z}(X(S) - \alpha Y(S)) = \sum_{z \in Z}[(X(z) - \alpha Y(z)]_{+},$$

where $\alpha \geq 1$ and $[x]_{+} = \max\{x, 0\}$. We observe that the following holds directly from the definition of the hockey stick divergence.

Corollary 2. *For any probability measures X, Y over Z and for any ϵ, δ, it holds*

$$X \approx_{\epsilon,\delta} Y \text{ if and only if } \mathsf{D}^{\mathsf{hs}}_{e^{\epsilon}}(X, Y) \leq \delta \text{ and } \mathsf{D}^{\mathsf{hs}}_{e^{\epsilon}}(Y, X) \leq \delta.$$

Proof of Lemma 6. Let $\eta_{-\theta} = 1/2 - \theta$ and $\eta_{+\theta} = 1/2 + \theta$. For brevity, we let C denote $\mathsf{PB}(n, p_J)$. For any distribution $\mathcal{D}$, let $P_{\mathcal{D}}$ denote the probability measure with respect to $\mathcal{D}$. We first show that for any $\epsilon > 0$, it holds that $\mathsf{D}^{\mathsf{hs}}_{e^{\epsilon}}(P_C, P_{C+1})$ is at most

$$\max\left(\mathsf{D}^{\mathsf{hs}}_{e^{\epsilon}}\left(P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{+\theta})}, P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{+\theta})+1}\right), \mathsf{D}^{\mathsf{hs}}_{e^{\epsilon}}\left(P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{-\theta})}, P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{-\theta})+1}\right)\right).$$

We start with an upper bound of the hockey-stick divergence is reached at extreme points. We rely on the results in [COK22]. Although they use the Renyi divergence, their results are general enough to be applied to any f-divergence.

Lemma 8. *([COK22, Lemma 3.5])*

$$\mathsf{D}^{\mathsf{hs}}_{e^{\epsilon}}(P_C, P_{C+1}) \leq \max_{j \in [n]} \mathsf{D}^{\mathsf{hs}}_{e^{\epsilon}}\left(P_{B(j,\eta_{-\theta})+B(n-j,\eta_{+\theta})}, P_{B(j,\eta_{-\theta})+B(n-j,\eta_{+\theta})+1}\right). \tag{1}$$

Next, we apply data processing inequality to simplify (1) from the above lemma.

Lemma 9. *([COK22, Lemma 3.6]) (1) is upper bounded by*

$$\max\left(\mathsf{D}^{\mathsf{hs}}_{e^{\epsilon}}\left(P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{+\theta})}, P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{+\theta})+1}\right), \mathsf{D}^{\mathsf{hs}}_{e^{\epsilon}}\left(P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{-\theta})}, P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{-\theta})+1}\right)\right). \tag{2}$$

We extend the above to upper bound the hockey-stick divergence between probability measures differed by an integer amount greater than 1, i.e., P_C and P_{C+s} for $s > 1$.

Corollary 3. *For any $\epsilon > 0$, it holds that $\mathsf{D}^{\mathsf{hs}}_{e^{\epsilon}}(P_C, P_{C+s})$ is at most*

$$\max\left(\mathsf{D}^{\mathsf{hs}}_{e^{\epsilon}}\left(P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{+\theta})}, P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{+\theta})+s}\right), \mathsf{D}^{\mathsf{hs}}_{e^{\epsilon}}\left(P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{-\theta})}, P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{-\theta})+s}\right)\right).$$

Finally, to give a bound on the divergence, we can apply Lemma 3 to argue that the binomial distribution hides the small sensitivity. Specifically, as $\lceil\frac{n}{2}\rceil \in \Theta(\kappa)$ and $s = \lg\lg\kappa$, we can claim (ϵ, δ)-DDP with $\delta = \mathsf{negl}(\kappa)$.

Similarly, it holds that $\mathsf{D}^{\mathsf{hs}}_{e^{\epsilon}}(P_{C+s}, P_C) \leq \mathsf{negl}(\kappa)$. □

E Proof of Theorem 4

On the definition of a θ-good iteration. We keep the same definition of a θ-good iteration, except we set the exponent to $1/n'_R$, instead of $1/n_R$, and we also require $\theta \leq 1/10$. In particular,

- $\min h(A) = \min h(I)$ and $\min h(I) \in \left[1 - \left(\frac{1}{2} + \theta\right)^t, 1 - \left(\frac{1}{2} - \theta\right)^t\right]$ with $\boxed{t = \dfrac{1}{n'_R}}$.

Bundle of good iterations K_θ. The total number of iterations in the min-hash protocol π_{NMH} is $k = \Omega(\kappa \cdot \lg\lg\kappa)$. We require that $n_R/k^2 = \Omega(\kappa)$.

Using Lemma 4, with all but negligible probability, at least $\Omega(\kappa \cdot \lg\lg\kappa)$ iterations are θ-good. Recall that G_θ denotes the set of θ-good iterations, and $K_\theta = G_\theta \setminus S_{x^*}$. We set $k_g = |K_\theta|$. We further divide these k_g iterations into $u = \lg\lg\kappa$ bundles, each of which is of size $k_b = \Omega(\kappa)$. Those bundles are denoted by $K_{\theta,1}, \ldots, K_{\theta,u}$. We also let $K_{bad} := \overline{K_\theta}$.

Random variables for the protocol output. Let out^+_{bad} be the protocol's match count for sets A, B_{+x^*} w.r.t. the hash functions in K_{bad}:

$$out^+_{bad} := |\{j \in K_{bad} : \min h_j(A) = \min h_j(B_{+x^*})\}|\,.$$

Likewise, let out_{bad} be the number of matches for sets A and B (instead of B_{+x^*}) in iterations in K_{bad}. Similarly, for $i \in [u]$, we let out^+_i and out_i denote the output for the i-th bundle, with or without x^* respectively. Note that $out^+_i = out_i$, since we ruled out S_{x^*} from K_θ. Note that the final output of the min-hash protocol for input B_{+x^*} is equal to $out^+_{bad} + \sum_{i=1}^u out_i$; the final output for input B is $out_{bad} + \sum_{i=1}^u out_i$. Let

$$\vec{out} = out^+_{bad} || out_{bad} || out_1 || \cdots || out_u.$$

We also consider the output with the ith bundle missing; that is, for $i \in [u]$ let

$$\vec{out}_{-i} = out^+_{bad} || out_{bad} || out_1 || \cdots || out_{i-1} || out_{i+1} || \cdots || out_u.$$

Upper-bounding leakage from the output. Since $|K_{bad}|$ and $|K_{\theta,i}|$ are at most $k \in poly(\kappa)$, we can safely assume that the total number of bits in $\vec{out}$ is

$$2\lg|K_{bad}| + \sum_{i=1}^{u} \lg|K_{\theta,i}| \le (2 + \lg\lg\kappa)\lg|poly(\kappa)| \le \kappa.$$

Distribution of R and its min-entropy. The original distribution on the secret set R is the uniform distribution over all sets of size n_R with each element is chosen from a universe $\mathcal{U}$. The universe $\mathcal{U}$ has size $\ell \cdot n_R$ with $\ell \ge 4(n_R)^3$.

Now choose, uniformly at random, a partition $\{U_1, \ldots, U_{n_R}\}$ of $\mathcal{U}$ where each $|U_j| = \ell$ such that the element in the jth slot of R belongs to U_j. These universes $\{U_1, \ldots, U_{n_R}\}$ are leaked in the analysis.

Let $\mathcal{D}$ denote the original distribution over the set R, but conditioned on the leaked information $\{U_1, \ldots, U_{n_R}\}$. The distribution $\mathcal{D}$ is equivalent to a distribution over streams of n_R elements, where the element in the i-th slot is chosen uniformly at random from U_i. Therefore, $\mathcal{D}$ has min-entropy $n_R \lg\ell$.

We additionally consider arbitrary leakage $f(R)$ of length L such that

$$n_R \lg\ell - L \ge \frac{8n_R}{9}\lg\ell + 2n_R.$$

Available iterations in a bundle. For a fixed set $Z \subseteq R$, in a min-hash graph, we say that a set of iterations in the ith bundle $K_{\theta,i}$ is available with respect to Z if there are no edges from Z to that set. In other words, no elements in Z contribute to the final count reduction for any of those iterations. In this sense, those iterations are is still available for the count reduction by the other elements than those in Z. More formally, consider a graph $G \leftarrow \mathbf{MinhashG}_{H_1}(A, I, x^*, H_2)$ and letting $G = (\mathcal{X}, \mathcal{Y}, \mathcal{E})$, we define

$$\mathsf{Avail}_G(K_{\theta,i}, Z) := \{j \in K_{\theta,i} : \forall z \in Z : (z, j) \notin \mathcal{E})\}.$$

Existence of a good bundle. We now describe an experiment to check if the ith bundle of iterations is good in the sense that given the fixed hash, the distribution $\mathcal{D}$ (after the leakage) satisfies the DP-like property conditions specified in Lemma 7. Roughly speaking, Lemma 7 shows that a bundle will be good with a high probability.

Process $\mathbf{IsAGoodBundle}(i, \vec{out}_{-i}, \mathcal{D}, A, I, x^*, H_1, H_2)$:

1. Consider $G \leftarrow \mathbf{MinhashG}_{H_1}(A, I, x^*, H_2)$.
2. Let $\mathcal{D}_{1,i} := \mathcal{D} \mid \vec{out}_{-i}$. In other words, $\mathcal{D}_{1,i}$ is the distribution $\mathcal{D}$ on R, but conditioned on the output vector $\vec{out}_{-i}$. If $\mathcal{D}_{1,i}$ has min-entropy less than $n_R \lg(\ell) - L - 2\kappa$ then output $\mathsf{FAIL}_{1,i}$ and terminate.
3. Check if if there is a leakage function $f_G(R)$ which leaks $V = \{j_1, \ldots, j_{n'_R}\}$ and $T = \mathsf{Avail}_G(K_{\theta,i}, R \setminus R')$ such that there exists a distribution with the Geometric Collision Property over sets $R' = \{x_j \in R : j \in V\}$. If there is no such distribution, output $\mathsf{FAIL}_{2,i}$ and terminate. Let $\mathcal{D}_{2,i} := \mathcal{D}_{1,i}|f_G(R)$.
4. If it holds $|T| \le \frac{1}{10}|K_{\theta,i}|$, output $\mathsf{FAIL}_{3,i}$ and terminate. Let $k_v = |T|$.
5. Compute $D_{T,r}(\mathcal{D}_{2,i})$ and check if $D_{T,r}$ satisfies the conditions given in Lemma 7. Output $\mathsf{FAIL}_{4,i}$ and terminate, if the above check fails.
6. Output SUCCESS.

Failure probability $\mathsf{FAIL}_{1,i}$. We claim that $\mathsf{FAIL}_{1,i}$ takes place with a negligible probability. By applying [DORS08, Lemma 2.2], the average min-entropy of $\mathcal{D}|\vec{out}_{-i}$ is at least $n_R \lg \ell - L - \kappa$, which implies that the min-entropy of $\mathcal{D}|\vec{out}_{-i}$ is at least $n_R \lg \ell - L - 2\kappa \ge \frac{8n_R}{9} \lg \ell + n_R$ with probability $1 - 2^{-\kappa}$ (assuming that $n_R \ge 2\kappa$).

Failure probability $\mathsf{FAIL}_{2,i}$.

Lemma 10. *The experiment outptus $\mathsf{FAIL}_{2,i}$ with a negligible probability.*

We give the proof later in Appendix H.

Failure probability $\mathsf{FAIL}_{3,i}$ We show that $\mathsf{FAIL}_{3,i}$ occurs with negligible probability. Let $n = n_R$ and $n' = n'_R$ for brevity of notation. Recall that $n' = n/3$. Let X_j be an indicator variable that represents whether there is an edge from $(n - n')$ nodes to iteration j. Therefore, we have

$$\Pr_{H_2}[|T| = r] = \Pr_{H_2}\left[\sum_{j=1}^{k_b} X_j = k_b - r\right].$$

Recall that $p_j \le 1 - (\eta_{-\theta})^{1/n'}$ and $\Pr[X_j = 1] = 1 - (1 - p_j)^{n-n'} \le 1 - (\eta_{-\theta})^{\frac{n-n'}{n'}} = 1 - (\eta_{-\theta})^2 \le 1 - (2/5)^2$. Therefore, we have

$$m := \mathbf{E}\left[\sum_{j=1}^{k_b} X_j\right] \le k_b \cdot (1 - (\eta_{-\theta})^2) \le 0.84 k_b$$

Using the Chernoff bound and due to $k_b \in \Omega(\kappa)$, we have

$$\Pr_{H_2}\left[|T| \le \frac{k_b}{10}\right] = \Pr_{H_2}\left[\sum_{j=1}^{k_b} X_j \ge \frac{9}{10}k_b\right] \le \exp\left(-\frac{(0.9k_b - m_0)^2}{2m_0}\right) = \exp(-\Omega(\kappa)).$$

Failure probability $\mathsf{FAIL}_{4,i}$. By Lemma 7, for all $i \in [u]$, conditioned on $\mathsf{FAIL}_{1,i}$, $\mathsf{FAIL}_{2,i}$, $\mathsf{FAIL}_{3,i}$ not occurring, let

$$p_4 := \Pr_{H_1,H_2}[\mathbf{IsAGoodBundle}(i, \vec{out}_{-i}, \mathcal{D}, A, I, x^*, H_1, H_2) = \mathsf{FAIL}_{4,i}].$$

Then, we have $p_4 \in O(k_v \lg^3(\kappa)/(n_R)^{0.5})$.

Existence of a good bundle out of u bundles. Observe that conditioned on $\mathsf{FAIL}_{1,i}$, $\mathsf{FAIL}_{2,i}$, $\mathsf{FAIL}_{3,i}$ not occurring, the process outputs $\mathsf{FAIL}_{4,i}$ independently of $(\vec{out}_{-i}, R')$, since the hash values in H_2 for any iteration are chosen independently of those for the other iterations. Using the above, since $k_v/\sqrt{n_R} = O(1/\sqrt{\kappa})$, we have the following:

> *The experiment* **IsAGoodBundle** *outputs SUCCESS for at least one bundle with probability* $1 - p_4^u = 1 - \mathsf{negl}(\kappa)$.

Noise distribution. We define a noise distribution Φ and give an analysis of the hockey stick divergence of $\Phi(r)$ and $\Phi(r - \lg\lg(\kappa))$.

Definition 10 (Noise distribution Φ)**.** We define $\Phi(r)$ as follows:

- Choose H_1 and H_2 randomly.
- Let $i^* \in [u]$ be the index to the bundle that **IsAGoodBundle** outputs SUCCESS.
- For $r \in [0, k_v]$, output $D_{T,r}(\mathcal{D}_{2,i^*})$, where $T = \mathsf{Avail}_G(K_{\theta,i^*}, R \setminus R')$.
- For $r \notin [0, k_v]$, $\Phi(r) := 0$

Lemma 11. *The hockey stick divergences* $\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}(\Phi(r), \Phi(r - \lg\lg(\kappa)))$ *and* $\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}(\Phi(r - \lg\lg(\kappa)), \Phi(r))$ *are both negligible in* κ*.*

Proof. For brevity, for any r, denote $D_r := D_{T,r}(\mathcal{D}_{2,i^*})$. Conditioned on **IsAGoodBundle** outputting SUCCESS with input $\vec{out}_{-i^*}$, we have a and b such that for $r \in [a + \lg\lg\kappa, b]$,

$$e^{-\epsilon} \le \frac{e^{-\epsilon/3} E^{n'}_{k_v,r}}{e^{\epsilon/3} E^{n'}_{k_v,r-\lg\lg(\kappa)}} \le \frac{D_r}{D_{r-\lg\lg(\kappa)}} \le \frac{e^{\epsilon/3} E^{n'}_{k_v,r}}{e^{-\epsilon/3} E^{n'}_{k_v,r-\lg\lg(\kappa)}} \le e^{\epsilon}.$$

The first and last inequalities are from Corollary 1. The second and third inequalities are from the condition that the process outputs SUCCESS. The hockey stick divergence $\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}(\Phi(r), \Phi(r - \lg\lg(\kappa)))$ is therefore at most

$$\sum_{r \notin [a+\lg\lg\kappa, b]} D_r \le k_v \cdot \mathsf{negl}(\kappa) = \mathsf{negl}(\kappa). \qquad \square$$

Similarly, $\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}(\Phi(r - \lg\lg(\kappa)), \Phi(r))$ is also $\mathsf{negl}(\kappa)$.

Putting it all together. Let c be the final count produced by running protocol π_{NMH}. We consider the probabilities

$$\Pr_{H_1,H_2,\mathcal{D}}[c \mid B_{+x^*}] \quad \text{and} \quad \Pr_{H_1,H_2,\mathcal{D}}[c \mid B].$$

We consider only runs of the protocol that yield c and for which there exists some $i^* \in [u]$ such that the process **IsAGoodBundle** returns SUCCESS given $\vec{out}_{-i^*}$ as input. We just have argued that such an i^* exists with all but negligible probability.

Further, we consider only runs of the protocol for which $out^+_{bad} - out_{bad} \le s = \lg\lg(\kappa)$. By Lemma 2, this also occurs with all but negligible probability, We will also leak $k_v = |\mathsf{Avail}(K_{\theta,i^*}, R \setminus R')|$.

Conditioned on the above events, by the definition of the distribution Φ, the value out_{i^*} contributes $(k_v - r)$ to the final count c with probability $p = \Phi(r)$. Recall that every iteration j in K_{θ,i^*} is good, which means $\min h_j(A) = \min h_j(I)$, potentially contributing to the output.

Therefore, assuming none of bad events occur (which happens with overwhelming probability), by applying Lemma 11, the probability that the ratio of probabilities of a certain output out for B_{+x^*} and B is not contained in $[e^{-\epsilon}, e^{\epsilon}]$ is $\mathsf{negl}(\kappa)$, and therefore we conclude that the protocol satisfies the DDP security.

F Proof of Lemma 7

When considering the probability of $D_{T,r}(\mathcal{D})$ and $I_{R',T,r}$ over the choice of H_2, the identity of T doesn't matter except for its size $k_b = |T|$. Therefore, in this case, we will simply use $D_{k_b,r}(\mathcal{D})$ and $I_{R',k_b,r}$ Moreover, when it is clear from the context, we will sometimes omit k_b and $\mathcal{D}$ and say $E_r^{R'} = E_r^{n'_R}$ and $I_{R',r} = I_{R',k_b,r}$, and $D_r = D_{k_b,r}(\mathcal{D})$.

We first show the following lemma holds.

Lemma 12. *Let $\mathcal{D}$ be a distribution over sets of size n'_R with geometric collision property. Fix H_1 and consider k_b, θ, a, b specified in Lemma 1 with the same requirements. Then, we have the following:*

Case 1: *If $r \notin [a+s, b]$, we have $\Pr_{H_2}[D_{k_b,r}(\mathcal{D}) \le \mathsf{negl}(\kappa)] \ge 1 - \mathsf{negl}(\kappa)$.*

Case 2: *If $r \in [a, b]$, then we have*

$$\Pr_{H_2}\left[e^{-\epsilon/3} E_{k_b,r}^{n'_R} \le D_{k_b,r}(\mathcal{D}) \le e^{\epsilon/3} E_{k_b,r}^{n'_R}\right] \ge 1 - (e^{\epsilon/3} - 1)^{-2} \cdot \frac{16 \lg^3(\kappa)}{\sqrt{n_R}}.$$

Then, Lemma 7 follows by taking a union bound over different cases of $r \in [k_b]$.

F.1 Proof of Lemma 12

We also define $\rho(R') := \Pr_{R' \sim \tilde{\mathcal{D}}}[R']$.

Proof for Case 1. We first consider Case (1). By applying the Case (1) of Corollary 1, we have $E_r^{n'_R} \in \mathsf{negl}(\kappa)$. Given $E_r^{n'_R} \in \mathsf{negl}(\kappa)$, we show

$$\Pr_{H_2}[D_r(\mathcal{D}) \le \mathsf{negl}(\kappa)] \ge 1 - \mathsf{negl}(\kappa).$$

Recall that $D_r(\mathcal{D}) = \sum_{R'} \rho(R') \cdot I_{R',r}$. Assume toward the contradiction that the negation of the statement holds. This means there are polynomials p and q, and a collection Heavy of R's such that

$$\Pr_{H_2}\left[\sum_{R' \in \mathsf{Heavy}} \rho(R') \cdot I_{R',r} \ge 1/p(\kappa)\right] \ge 1/q(\kappa).$$

The above implies that $\sum_{R' \in \mathsf{Heavy}} \rho(R') \ge 1/p(\kappa)$. Now, since $\mathcal{D}$ and H_2 are independent, we have $\sum_{R' \in \mathsf{Heavy}} \rho(R') \Pr_{H_2}[I_{R',r}] \ge \frac{1}{p(\kappa)q(\kappa)}$. However, considering that $\Pr_{H_2}[I_{R',r}] = E_r^{n'_R}$, which is negligible, the above is a contradiction.

Proof for Case 2. We will bound $D_r = \sum_{R'} \rho(R') \cdot I_{R',r}$ using Chebyshev inequality. For this, we would like to bound the variance of D_r.

We start with showing the following lemma, which will allow us to ignore the tail when we bound the variance. Below, the value z will correspond to the size of the intersection of the two sets R'_i and R'_j.

Lemma 13. *Fix H_1. Consider a graph $G \leftarrow \mathbf{MinhashG}_{H_1}(A, I, x^*, H_2)$. Consider any set T of iterations in G such that $|T| = k_b$. Let Z be a set of left nodes in G such that $|Z| \le n'_R$. Let $z = |Z|$. Consider the probability (over the choice of H_2) that Z has more than $z \lg\lg\kappa$ outgoing edges in G. This probability is negligible in κ.*

Proof. Let $p = 1 - (\eta_{-\theta})^{1/n'_R}$. We first show that $p \leq 1/n'_R$. Recall $\theta \leq 1/10$, which implies $e^{-1} \leq 1/2 - \theta = \eta_{-\theta}$. Therefore, we have $(1 - 1/n'_R)^{n'_R} \leq e^{-1} \leq \eta_{-\theta}$, so $1 - 1/n'_R \leq (\eta_{-\theta})^{1/n'_R}$. Therefore, we have $p = 1 - (\eta_{-\theta})^{1/n'_R} \leq 1/n'_R$.

Let $\mathsf{Edges}(Z, T)$ be the set of edges from Z to T. Over the choice of H_2, the probability that each pair in $Z \times T$ forms an edge is at most p. Therefore, we can simply use a Binomial distribution to bound the probability. In particular, with $t = \lg\lg\kappa$ we have

$$\Pr_{H_2}\left[|\mathsf{Edges}(Z,T)| \geq zt\right] \leq \Pr\left[\mathrm{B}(z\hat{k}, p) \geq zt\right] \leq \binom{z\hat{k}}{zt} \cdot p^{zt} \leq \binom{z\hat{k}}{zt} \cdot \left(\frac{1}{n'_R}\right)^{zt} \leq \left(\frac{e\hat{k}}{tn'_R}\right)^{zt}$$

Since n'_R is much larger than k_b, the above probability becomes negligible in κ. □

Now we prove the following lemma towards bounding the variance of D_r.

Lemma 14. *Fix H_1. We set the parameters for k_b, a and b as stated in Lemma 7. Let R'_i, R'_j be sets of nodes on the left of size n'_R such that with $|R'_i \cap R'_j| = z$. Let $\zeta = z \lg\lg\kappa$. Then for all $a \leq r \leq b$, we have*

$$\Pr_{H_2}[I_{R'_i,r} \wedge I_{R'_j,r}] = \mathop{\mathbb{E}}_{H_2}[I_{R'_i,r} \cdot I_{R'_j,r}] \leq \left(1 + \frac{\zeta \cdot (e^{\zeta\epsilon/3} + 1)}{{\eta_{-\theta}}^{zk_b/n'_R}}\right)\left(E_r^{n'_R}\right)^2$$

Proof. Fix R'_i, R'_j with $|R'_i \cap R'_j| = z$. Let $Z = R'_i \cap R'_j$ and $X = R'_i - Z$. Then, we have

$$\begin{aligned}
\Pr_{H_2}[I_{R'_i,r} \wedge I_{R'_j,r}] &= \sum_{m=0}^{r} \Pr[I_{X,m} \wedge I_{Z,r-m} \wedge I_{R'_j,r}] \\
&\leq \sum_{m=0}^{r-\zeta} \Pr[I_{Z,r-m}] + \sum_{m=r-\zeta+1}^{r} \Pr[I_{X,m} \wedge I_{R'_j,r}] \\
&= \sum_{m=\zeta}^{r} \Pr[I_{Z,m}] + \sum_{m=r-\zeta+1}^{r} \Pr[I_{X,m}] \cdot \Pr[I_{R'_j,r}] \\
&\leq \mathsf{negl}(\kappa) + \sum_{m=r-\zeta+1}^{r} \Pr[I_{X,m}] \cdot \Pr[I_{R'_j,r}] \\
&= \mathsf{negl}(\kappa) + E_r^{n'_R} \cdot \sum_{m=r-\zeta+1}^{r} \Pr[I_{X,m}].
\end{aligned}$$

The second inequality holds due to Lemma 13.

It is left to bound $\Pr[I_{X,m}]$ for $m \in (r-\zeta, r]$. We observe that $\Pr_{H_2}[I_{X,m}] = \Pr[I_{R'_i,m} | \ I_{Z,0}]$. In other words, the event that X contributes to noise pattern m is equivalent to the event that R'_i contributes to m conditioned on the intersection having no contribution. Therefore, we have

$$\Pr_{H_2}[I_{X,m}] = \frac{\Pr[I_{R'_i,m} \wedge I_{Z,0}]}{\Pr[I_{Z,0}]} \leq \frac{\Pr[I_{R'_i,m}]}{{\eta_{-\theta}}^{zk_b/n'_R}} = \frac{E_m^{n'_R}}{{\eta_{-\theta}}^{zk_b/n'_R}}.$$

We now bound $E_m^{n'_R}$ for $m \in (r-\zeta, r]$. Let $m^* := \arg\max_m\{E_m^{n'_R} : m \in (r-\zeta, r]\}$.

Using Corollary 1 we have $E_{m^*}^{n'_R} \leq (e^{\epsilon/3})^{\zeta} \cdot E_r^{n'_R} + \mathsf{negl}(\kappa)$. Therefore, we have

$$\begin{aligned}
\Pr_{H_2}[I_{R'_i,r} \wedge I_{R'_j,r}] &\leq \mathsf{negl}(\kappa) + E_r^{n'_R} \cdot \sum_{m=r-\zeta+1}^{r} \Pr[I_{X,m}] \ \leq \ \mathsf{negl}(\kappa) + \zeta \cdot E_r^{n'_R} \cdot \Pr[I_{X,m^*}] \\
&= \mathsf{negl}(\kappa) + \zeta \cdot E_r^{n'_R} \cdot \frac{E_{m^*}^{n'_R}}{\eta_{-\theta}{}^{zk_b/n'_R}} = \mathsf{negl}(\kappa) + \zeta \cdot E_r^{n'_R} \cdot \frac{e^{\zeta\epsilon/3} E_r^{n'_R} + \mathsf{negl}(\kappa)}{\eta_{-\theta}{}^{zk_b/n'_R}} \\
&\leq \left(1 + \frac{\zeta \cdot (e^{\zeta\epsilon/3} + 1)}{\eta_{-\theta}{}^{zk_b/n'_R}}\right) \left(E_r^{n'_R}\right)^2
\end{aligned}$$

□

We set the parameters for H_1, k_b, a and b as stated in Lemma 7. Let $\mathcal{D}$ be a distribution with the geometric collision property. Then, we show that for every $a \leq r \leq b$, we have

$$\mathsf{Var}_{H_2}[D_r] \leq \frac{16 \lg^3(\kappa)}{\sqrt{n_R}} \left(E_{k_b,r}^{n'_R}\right)^2.$$

Consider any $r \in [a, b]$. Recall that $D_r := \sum_{R' \in \mathsf{Supp}(\tilde{\mathcal{D}})} \rho(R') \cdot I_{R',r}$.

$$\begin{aligned}
\mathsf{Var}_{H_2}[D_r] \quad &= \sum_{R'_i, R'_j} \rho(R'_i) \cdot \rho(R'_j) \cdot (\mathbb{E}[I_{R'_i,r} \cdot I_{R'_j,r}] - \mathbb{E}[I_{R'_i,r}] \cdot \mathbb{E}[I_{R'_j,r}]) \\
&\leq \sum_{R'_i, R'_j : |R'_i \cap R'_j| \geq 1} \rho(R'_i) \cdot \rho(R'_j) \cdot \mathbb{E}[I_{R'_i,r} \cdot I_{R'_j,r}] \\
&= \sum_{z=1}^{n'_R} \Pr_{R'_i, R'_j \sim \mathcal{D}}[|R'_i \cap R'_j| = z] \cdot \mathbb{E}[I_{R'_i,r} \cdot I_{R'_j,r}] \\
&\leq \sum_{z=1}^{n'_R} \left(\frac{1}{\sqrt{n_R}}\right)^z \cdot \left(1 + \frac{\zeta \cdot (e^{\zeta\epsilon/3} + 1)}{\eta_{-\theta}{}^{zk/n'_R}}\right) \cdot \left(E_r^{n'_R}\right)^2 \\
&\leq \sum_{z=1}^{n'_R} \left(\frac{1}{\sqrt{n_R}}\right)^z \cdot \left(\zeta \cdot \frac{e^{\zeta} + 2}{(2/5)^{\zeta/3}}\right) \cdot \left(E_r^{n'_R}\right)^2 \\
&\leq \sum_{z=1}^{n'_R} \left(\frac{1}{\sqrt{n_R}}\right)^z \cdot (8^{\zeta+1}) \cdot \left(E_r^{n'_R}\right)^2
\end{aligned}$$

The first inequality holds because if R'_i are R'_j are disjoint, then $I_{R'_i,r}$ and $I_{R'_j,r}$ are independent over the choice of H_2, and the relevant terms are canceled out. The second inequality is due to the geometric collision property of $\mathcal{D}$ and Lemma 14. The third inequality holds with $\epsilon \leq 3$ since $\theta < 1/10$ and k_b is much smaller than n'_R. Therefore, we have $\mathsf{Var}_{H_2}[D_r] \leq 8 \cdot \left(E_r^{n'_R}\right)^2 \cdot \sum_{z=1}^{n'_R} \left(\frac{\lg^3 \kappa}{\sqrt{n_R}}\right)^z \leq \frac{16 \lg^3 \kappa}{\sqrt{n_R}} \left(E_r^{n'_R}\right)^2$. □

Finally, by Chebyshev, we have that for all $a \leq r \leq b$,

$$\begin{aligned}
\Pr_{H_2}\left[D_r \notin [e^{-\epsilon/3}(E_{k_b,r}^{n'_R}), e^{\epsilon/3}(E_{k_b,r}^{n'_R})]\right] &\leq \Pr\left[|D_r - E_{k_b,r}^{n'_R}| \geq (1 - e^{-\epsilon/3}) \cdot E_{k_b,r}^{n'_R}\right] \\
&\leq \frac{\mathsf{Var}[D_r]}{(1 - e^{-\epsilon/3})^2 \cdot (E_{k_b,r}^{n'_R})^2} \leq \frac{16 \lg^3(\kappa)}{(1 - e^{-\epsilon/3})^2 \sqrt{n_R}}.
\end{aligned}$$

G Strong Chain Rule

Strong chain rule for a special case: achieving flatness through clustering. Fortunately, a stronger version of the chain rule is known to hold for a special leakage pattern, i.e., when elements are conditioned *in order* [Skó19]; very roughly speaking, for every i, the min-entropy of $R_i|(R_1,\ldots,R_{i-1})$ is essentially the same as the min-entropy of $(R_1,\ldots,R_i)$ minus the min-entropy of $(R_1,\ldots,R_{i-1})$ at the sacrifice of an additional small leakage, which is called a *spoiling leakage*.

They achieve this by grouping possible sequences with a similar distributional characteristic into the same cluster. Then, in every cluster, the distribution of sequences conditioned on that cluster will be essentially flat. Now, the spoiling leakage corresponds to the cluster identifier. By making every cluster contain sufficiently many sequences (leading to sufficient min-entropy due to flatness), the total number of clusters can be small (leading to a short spoiling leakage).

Notes on notations. For brevity, in this section, we omit the subscript from n_R, i.e., we denote $n = n_R$. For any sequence of random variables $R = R_1,\ldots,R_n$ (for the secret input R), we denote $R_{<i} = R_1,\ldots,R_{i-1}$ and $R_{\leq i} = R_1,\ldots,R_i$. Likewise, we extend such subscript notations and use $R_{>i}$ and $R_{\geq i}$. We use lower case $r = r_1,\ldots,r_n$ to denote the actual set/sequence.

Strong chain rule for our setting. We first adapt the result in [Skó19] into our setting. Then, we argue that a sufficient number of elements still have high min-entropy, even conditioned on the previous elements. Finally, we show that these high min-entropy (conditioned) elements provide the geometric collision property.

Theorem 6 (Block structures with few bits spoiled in our setting)**.** *We consider a min-hash graph $G = (\mathcal{X},\mathcal{Y},\mathcal{E})$ constructed from $\mathbf{MinhashG}_{H_1}(A, I, x^*, H_2)$, while focusing on a single bundle $K_{\theta,*}$ of iterations.*

Let $\mathcal{U} = U_1 \times \cdots \times U_n$ be a fixed universe and $R = (R_1,\ldots,R_n)$ be a sequence of (possibly correlated) random variables where each R_i is over U_i (and all are disjoint) and $|U_i| = \ell$ for all i. Then, for any $\epsilon \in (0,1)$ and any $\delta > 0$, there exists a spoiling leakage function $f_G(R)$ that satisfies the following properties.

1. *It holds that $\Pr_R[f(R) = \perp] \leq \epsilon n$.*
2. *Let $Im(f)$ be the set of images of f. Every $y \in Im(f) \setminus \{\perp\}$ specifies two disjoint sets V and W such that $V \cup W = [n]$.*
3. *Conditioned on any $y \in Im(f) \setminus \{\perp\}$, for every $i \in V$, every element in distribution $R_i \mid R_{<i}$ has low probability weight, i.e.,*
$$\forall y \in Im(f) \setminus \{\perp\}, \forall r \text{ s.t. } f(r) = y, \forall i \in V: \quad \Pr\left[R_i = r_i \;\middle|\; R_{<i} = r_{<i},\ y\right] \leq \frac{2^\delta}{n^{1.5}}.$$
4. *Conditioned on any $y \in Im(f) \setminus \{\perp\}$, for every $i \in W$, it holds that $R_i \mid R_{<i}$ has small support size, i.e.,*
$$\forall y \in Im(f) \setminus \{\perp\}, \forall r \text{ s.t. } f(r) = y, \forall i \in W:$$
$$|\{r_i : \Pr[R_i = r_i | R_{<i} = r_{<i}, y]] \geq 0\}| \leq 2^\delta \cdot n^{1.5}.$$
5. $|Im(f)| \leq n \cdot (2e)^{n/2} \cdot \frac{(n+k_b)!}{n!} \cdot (2(\lg(\ell) + \lg(1/\epsilon))/\delta)^n$.
6. $\mathsf{Avail}_G(K_{\theta,*}, R_W)$ *can be computed from $f(R)$, where $R_W := \{R_i : i \in W\}$.*

G.1 Proof of Theorem 6

By following the general idea of [Skó19], we will build clusters, and the spoiling leakage will be the cluster identifier. However, we will slightly change the way we build clusters.

Condition 1. Throughout our proof, we let $\Pr[r_i]$ denote $\Pr[R_i = r_i]$ for brevity, whenever the referred random variable is clear. Before forming the clusters, we will first like to exclude all sequences $r \in \mathcal{U} = U_1 \times \cdots \times U_n$ having a very small probability $\Pr_R[R_i = r_i \mid R_{<i} = r_{<i}] < \epsilon/\ell$ for any $i \in [n]$ and only consider the remaining $\mathcal{U}' \subset \mathcal{U}$. Specifically, we let $f(r) = \perp$ for all $r \notin \mathcal{U}'$. As we will see later, this probability lower bound is vital to upper bound $|Im(f)|$.

Claim. *Let $\mathcal{U}'$ be the set containing all the sequences r such that $\Pr_R[R_i = r_i \mid R_{<i} = r_{<i}] \geq \epsilon/\ell$ for all $i \in [n]$. Then, we have $\Pr[r \in \mathcal{U}'] \geq 1 - \epsilon n$.*

Proof. For each $i \in [n]$, and any $r_{<i} \in U_1 \times \cdots \times U_{i-1}$, we have

$$\sum_{u \in U_i : \Pr_R[R_i = u \mid R_{<i} = r_{<i}] < \epsilon/\ell} \Pr_R[R_i = u \mid R_{<i} = r_{<i}] < \sum_{u \in U_i} \epsilon/\ell = \epsilon.$$

Therefore, using a union bound across all $i \in [n]$, we have $\Pr[r \notin \mathcal{U}'] \leq \epsilon \cdot n$. □

Building clusters. For each $r \in \mathcal{U}'$, we describe how to compute $f(r) = (f_1(r), f_2(r), \ldots, f_n(r))$, which will serve as the cluster identifier. Let $\mathsf{r}(a)$ denote a rounding function that rounds a to the closest multiple of $\delta/2$. We say $a \approx_{\mathsf{r}} a'$ if $\mathsf{r}(a) = \mathsf{r}(a')$.

For each r, do the following:

1. Let $f_{>n}(r) = \perp$ for any r, and initialize $W = \emptyset$.
2. For $i = n, \ldots, 1$, do the following:
   (a) Let $\text{SP}^1_i(r)$ denote the surprise of the ith element of r. More formally,

   $$\text{SP}^1_i(r) = -\lg \Pr_R[R_i = r_i \mid R_{<i} = r_{<i},\ f_{>i}(R) = f_{>i}(r)].$$

   This surprise measure represents how rare and surprising the event r_i is, conditioned on $r_{<i}$, $f_{>i}(r)$. In a sense, we will group sequences with similar surprises into a cluster.

   (b) Let $\text{SP}^2_i(r)$ denote the surprise of the sequences with a similar surprise level in aggregate.

   $$\text{SP}^2_i(r) = -\lg \Pr_R[\text{SP}^1_i(R) \approx_{\mathsf{r}} \text{SP}^1_i(r) \mid R_{<i} = r_{<i},\ f_{>i}(R) = f_{>i}(r)].$$

   Note $\text{SP}^1_i(r) \geq \text{SP}^2_i(r)$, since at least sequence r has $\text{SP}^1_i(r)$ and possibly more points may approximately share the surprise. Note also that $\text{SP}^2_i(r)$ is a deterministic function of $\text{SP}^1_i(r), r_{<i}, f_{>i}(r)$.

   (c) If $\mathsf{r}(\text{SP}^1_i(r)) - \mathsf{r}(\text{SP}^2_i(r)) \geq 1.5 \lg(n)$ then let $f_i(r) = (\mathsf{r}(\text{SP}^1_i(r)), \mathsf{true})$.

   (d) Otherwise, let $f_i(r) = (\mathsf{r}(\text{SP}^1_i(r)), \mathsf{false}, H_i)$ and add i to W. Here, H_i is defined as $N(\{r_i\}) \setminus N(r_W)$, where N refers to the neighbors (restricted to $K_{\theta,*}$) of the input set of nodes in G. In other words, H_i contains the iterations newly covered by element r_i; any iterations previously covered by r_W are ruled out in H_i. In this way, we can reduce the length of the cluster identifier.
3. Set $f(r) = f_1(r), \ldots, f_n(r)$. Set $V = [n] \setminus W$.

Conditions 2 and 3. Condition 2 follows from how V is computed in step 3. We now show that condition 3 holds. In particular, $\forall y \in Im(f(\cdot)) \setminus \{\perp\}, \forall r$ s.t. $f(r) = y, \forall i \in V$ we have

$$\Pr[r_i \mid r_{<i}, y] = \Pr[r_i \mid r_{<i}, y_{\geq i}] = \frac{\Pr[r_i \wedge r_{<i} \wedge y_{\geq i}]}{\Pr[r_{<i} \wedge y_{\geq i}]} = \frac{\Pr[r_i \wedge r_{<i} \wedge y_{>i}]}{\Pr[r_{<i} \wedge y_{>i}] \Pr[y_i \mid r_{<i} \wedge y_{>i}]}$$

The first equality is due to $y_{<i}$ being a deterministic function of $r_{<i}$, $y_{\geq i}$. Similarly, the nominator of the final fraction is due to y_i being a deterministic function of $r_{\leq i}$, $y_{>i}$. Moreover, $y_{i,2}$ (i.e., true) can be deterministically computed from $y_{i,1}$(i.e., $\mathsf{r}(\textsc{sp}_i^1(r))$), $r_{<i}, y_{>i}$. Therefore, the above is equal to

$$\frac{\Pr[r_i \wedge r_{<i} \wedge y_{>i}]}{\Pr[y_{i,1} \mid r_{<i} \wedge y_{>i}] \Pr[r_{<i} \wedge y_{>i}]} = \frac{\Pr[r_i \mid r_{<i} \wedge y_{>i}]}{\Pr[y_{1,i} \mid r_{<i} \wedge y_{>i}]} = \frac{2^{-\textsc{sp}_i^1(r)}}{2^{-\textsc{sp}_i^2(r)}} \leq \frac{2^{\delta}}{n^{1.5}}. \tag{3}$$

The last inequality holds since $i \in V$, $\mathsf{r}(\textsc{sp}_i^1(r)) - \mathsf{r}(\textsc{sp}_i^2(r)) \geq 1.5 \lg(n)$.

Condition 4. For r, y, i as quantified in the theorem statement, we have

$$\begin{aligned}
&|\{r_i : \Pr[R_i = r_i | R_{<i} = r_{<i}, y]] \geq 0\}| \\
&= |\{r_i : \Pr[R_i = r_i \wedge R_{<i} = r_{<i} \wedge y]] \geq 0\}| \\
&= |\{r_i : \Pr[R_i = r_i \wedge R_{<i} = r_{<i} \wedge y_{i,1}, y_{i,2}, y_{>i}]] \geq 0\}| \\
&\leq |\{r_i : \Pr[R_i = r_i | R_{<i} = r_{<i}, y_{i,1}, y_{i,2}, y_{>i}]] \geq 0\}|
\end{aligned}$$

By a similar argument as above, for all r_i s.t. $\Pr[R_i = r_i | R_{<i} = r_{<i}, y_{i,1}, y_{i,2}, y_{>i}]] \geq 0$, it holds

$$\Pr[R_i = r_i | R_{<i} = r_{<i}, y_{i,1}, y_{i,2}, y_{>i}] = \frac{\Pr[r_i \mid r_{<i} \wedge y_{>i}]}{\Pr[y_{i,1} \mid r_{<i} \wedge y_{>i}]} = \frac{2^{-\textsc{sp}_i^1(r)}}{2^{-\textsc{sp}_1^2(r)}} \geq \frac{2^{-\delta}}{n^{1.5}}$$

where the inequality holds since $i \in W$, we know that $\mathsf{r}(\textsc{sp}_i^1(r)) - \mathsf{r}(\textsc{sp}_i^2(r)) \leq 1.5 \lg(n)$. This means that $|\{r_i : \Pr[R_i = r_i | R_{<i} = r_{<i}, y_{\geq i}]] \geq 0\}| \leq 2^{\delta} \cdot n^{1.5}$.

Condition 5. To bound $|Im(f)|$, we first upper bound $y_{i,1}$. Recall that $\Pr_R[R_i = r_i \mid R_{<i} = r_{<i}] \geq \epsilon/\ell$ for all $i \in [n]$ and $r \in \mathcal{U}'$. Therefore, $\Pr_R[R_i = r_i \mid R_{<i} = r_{<i}, y] \geq \epsilon/\ell$ for all $i \in [n]$, and $\forall r$ such that $f(r) = y$.

Therefore, for all $r \in \mathcal{U}', i \in [n]$, we have $\textsc{sp}_i^1(r) \leq \lg(\ell) + \lg(1/\epsilon)$, which implies that $y_{i,1}$ has at most $2(\lg(\ell)) + \lg(1/\epsilon))/\delta$ different possibilities.

To upper bound the number of possibilities of the remaining parts, it suffices to upper bound the number of choices for set W of size m, as well as the number of possibilities for H_i's in each slot $i \in W$. Clearly, the former is $\binom{n}{m}$. For the latter part, note that each iteration appears at most once over all m slots. Therefore, the problem becomes how we can assign k_b different iterations into $m+1$ positions (with some positions possibly containing none) while assigning them to the $m+1$th position when they never appear in any slot of W. This is a well-known problem of stars and bars with $m+1$ variables and sum k_b, which has $\binom{m+k_b}{m}$ possibilities. Since we have $k_b!$ different orderings for k_b iterations, the upper bound is $\binom{m+k_b}{m} \cdot (k_b!)$. We have:

$$\begin{aligned}
|Im(f)| &\leq (2(\lg(\ell) + \lg(1/\epsilon))/\delta)^n \left(\sum_{m=0}^{n} \binom{n}{m} \binom{m+k_b}{m} \cdot k_b! \right) \\
&= (2(\lg(\ell) + \lg(1/\epsilon))/\delta)^n \left(\sum_{m=0}^{n} \binom{n}{m} \frac{(m+k_b)!}{m!} \right) \\
&\leq n \cdot \binom{n}{n/2} \cdot \frac{(n+k_b)!}{n!} \cdot (2(\lg(\ell) + \lg(1/\epsilon))/\delta)^n \\
&\leq n \cdot (2e)^{n/2} \cdot \frac{(n+k_b)!}{n!} \cdot (2(\lg(\ell) + \lg(1/\epsilon))/\delta)^n .
\end{aligned}$$

Condition 6. Finally, condition 6 follows from the definition of the clustering procedure. In particular, $H_W = \bigcup_{i \in W} H_i$ contains all the iterations that r_W covers. The available set can be computed by $K_{\theta,*} \setminus H_W$. This concludes our proof.

G.2 Generalization

It can be seen that in the above proof, the only properties that we used of the additional leakage H_i is that for $i \in W$, H_i depends only on $R_i, y_{>i}$ and that the number of choices for the output of the sequence of leakages $[H_i]_{i \in W}$ is bounded by some B. Theorem 5 stated in Section 8.5 is restatement of Theorem 6 with respect to any such leakage function. Note that the leakage functions ℓ_i specified above can model leakage with respect to a random oracle h, by letting $\rho_i = h(R_i)$.

H Proof of Lemma 10

For brevity, we denote $\mathcal{D} = \mathcal{D}_{2,i}$ and $\mathcal{D}_{\mathsf{leak}} = \mathcal{D}_{1,i}$ in the experiment $\mathsf{IsAGoodBundle}$. We show that $\mathcal{D}$ has the geometric collision property. In other words, we would like to show that when R is chosen uniformly at random from universe $\mathcal{U}$ then the distribution of these $n_R = n_B - n_I$ elements has the geometric collision property even with the leakage.

Towards this goal, by applying Theorem 6 to this distribution, we show that even with the leakage, there are at least $n_R/3$ elements that preserves enough min-entropy. We next show how these elements with sufficient min-entropy give the geometric collision property.

Remark 1 (Getting rid of tiny parts). Similar to [Skó19, Remark 2], we can further require that each cluster should have a probability that is “not too small”. Therefore, we define a new leakage function f' by substituting the ϵ in the above theorem with $\epsilon/2$, and additionally letting $f'(r) = \perp$ for all r such that $y \in f(r)$ and $\Pr_R[f(R) = y] < \epsilon n/(2|Im(f)|)$ (their total probability is at most $\epsilon n/2$), we obtain the following: f' satisfies all conditions in Theorem 6. Additionally, $\forall y \in Im(f')$, we have $\Pr_R[f'(R) = y] \geq \epsilon n/(2|Im(f)|)$.

Fraction of blocks with high entropy. Using Theorem 6, with setting $\ell \geq 4n^3$ and assuming sufficient min-entropy of R, we first show that one can ensure more than 1/3 fraction of the blocks having min-entropy at least $1.5\lg(n)$, upon leaking the outcome of f' and all previous blocks.

First notice that with all but ϵn probability, $f'(R) \neq \perp$. Therefore, it suffices to let $\epsilon = 2^{-\kappa}$. Then, by setting $\delta = 1$, we have

$$\begin{aligned}
\lg(|Im(f)|) &\leq n \cdot (2e)^{n/2} \cdot (n + k_b)^{k_b} \cdot (2(\lg(\ell) + \lg(1/\epsilon))/\delta)^n \\
&= \Big(\lg(n) + n/2 \cdot \lg(2e)\Big) + k_b \cdot \lg(n + k_b) + n \cdot (1 + \lg(\lg(\ell) + \kappa)) \\
&< 3n/2 + 2k_b \lg n + n(2 + \lg \kappa) < 0.5 n \lg n
\end{aligned}$$

for sufficiently large n with $k_b = \Omega(\kappa)$ and $n/k_b^2 = \Omega(\kappa)$.

Combining the above with Remark 1, we have $\Pr_{\mathcal{D}_{\mathsf{leak}}}[f'(R) = y] \geq \epsilon n/(2 \cdot 2^{0.5n \lg(n)})$ and for every $y \in Im(f') \setminus \{\perp\}$. Moreover, for every r such that $f'(r) = y$, we have

$$\Pr_{\mathcal{D}}[r] = \Pr_{\mathcal{D}_{\mathsf{leak}}}[r \mid y] = \frac{\Pr_{\mathcal{D}_{\mathsf{leak}}}[r \wedge y]}{\Pr_{\mathcal{D}_{\mathsf{leak}}}[y]} \leq \frac{2^{-(\frac{8n}{9} \lg \ell + n)}}{(\epsilon n/2) \cdot 2^{-0.5n \lg n}} = 2^{-(\frac{8}{9} \log \ell - 0.5 \lg n + 1) \cdot n} \cdot (2/\epsilon n), \tag{4}$$

which suggests $\mathcal{D}$ has min-entropy at least $(\frac{8}{9} \log \ell - 0.5 \lg n + 1) \cdot n - \lg(2/\epsilon n)$. We show that the following holds: The min-entropy of at least $n' = n/3$ blocks, *conditioned on the outcome of all prior blocks* as well as y, is at least $\lg(n^{1.5})$.

Towards a contradiction, assume otherwise. Let V be the set of blocks with min-entropy at least $\lg(n^{1.5})$ and let W be the set of blocks with min-entropy less than $\lg(n^{1.5})$ (as defined in Theorem 6). We will show that if $|V| \leq n/3$ there exists a point r in the support of $\mathcal{D}$ such that $\Pr_{\mathcal{D}}[r] > 2^{-(\frac{8}{9} \log \ell - 0.5 \lg n + 1) \cdot n} \cdot (2/\epsilon n)$, contradicting the min-entropy of $\mathcal{D}$.

First, find any value r_V^* such that $\Pr_{\mathcal{D}}[R_V = r_V^*] \geq \frac{1}{\ell^{|V|}}$. Note that r_V^* must exist since the support size of R_V is at most $\ell^{|V|}$. Let $\mathsf{Supp}_W(r_V^*) = \{r : r_V = r_V^* \wedge \Pr_{\mathcal{D}}[R = r] > 0\}$. Then, we have $\Pr_{\mathcal{D}}[R \in \mathsf{Supp}_W(r_V^*)] = \Pr[R_V = r_V^*] \geq \frac{1}{\ell^{|V|}}$.

Second, we show that $|\mathsf{Supp}_W(r_V^*)| \leq (2 \cdot n^{1.5})^{|W|}$. Consider any $r \in \mathsf{Supp}_W(r_V^*)$. Applying the fourth condition of Theorem 6 with $\delta = 1$, condition on any $y \in Im(f') \setminus \{\perp\}$, for any $i \in W$ and any fixing of $R_{<i} = r_{<i}$, the number of elements in the support of $R_i \mid r_{<i}$ is at most $2 \cdot n^{1.5}$, which implies that $|\mathsf{Supp}_W(r_V^*)|$ must be at most $(2 \cdot n^{1.5})^{|W|}$, since the positions for V are fixed to r_V^*.

Based on the above two arguments, by the averaging argument, there must be some $r^* \in \mathsf{Supp}_W(r_V^*)$ for which $\Pr_{\mathcal{D}}[R = r^*] \geq \frac{1}{(\ell)^{|V|}} \cdot \frac{1}{(2 \cdot n^{1.5})^{|W|}}$. Therefore, we have

$$\begin{aligned} -\lg \Pr_{\mathcal{D}}[r^*] &= |V| \lg(\ell) + |W| \lg(2n^{1.5}) = |V| \lg \ell + |W| + 1.5(n - |V|) \lg n \\ &\leq n + |V| \lg(\ell/n) + 1.5n \lg n \leq n + n/3 \lg(\ell/n) + 1.5n \lg n \\ &= n + n/3 \lg(\ell) - 1/3n \lg n + 1.5n \lg n, \end{aligned}$$

where the second to last line follows assuming $|V| < n/3$.

To reach contradiction to (4), we require that

$$n + n/3 \lg(\ell) - 1/3n \lg n + 1.5n \lg n \leq \left(\frac{8}{9} \lg \ell - 0.5 \lg n + 1\right) \cdot n - \lg(2/\epsilon n).$$

The above is implied by $5/3n \lg n \leq 5/9n \lg \ell - \lg(2/\epsilon n)$.

When $\ell \geq 4n^3$ the above is implied by $5/3n \lg n \leq 5/3n \lg n + 10/3n - \lg(2/\epsilon n)$, which is true for $n \geq \lg(1/\epsilon) = \kappa$. Thus we reach contradiction to (4). We therefore conclude that $|V| \geq n/3$.

Geometric collision property. Note that we can equivalently view R' in the support of $\mathcal{D}$ as a set of size n', or as a stream of elements of length n', where the element in the i-th block (for $i \in [n']$) comes from universe U_i, and $\{U_1, \ldots, U_{n'}\}$ are mutually disjoint. Taking the second view, given R', S' in the support of $\mathcal{D}$, we have that $|R' \cap S'| = z$ if and only if there exists some set $Z \subseteq [n']$ of size z such that (1) the *ordered* set of elements in the blocks of R' indexed by Z (denoted R'_Z) is equal to the *ordered* set of elements in the blocks of S' indexed by Z (denoted S'_Z) and (2) the set of elements in the blocks of R' indexed by $[n'] \setminus Z$ (denoted $R'_{\overline{Z}}$) and the set of elements in the blocks of S' indexed by $[n'] \setminus Z$ (denoted $S'_{\overline{Z}}$) are disjoint.

We are now ready to analyze the probability that $|R' \cap S'| = z$ for R', S' drawn from $\mathcal{D}$, and for $z \in [n']$:

$$\begin{aligned} \Pr_{R',S' \leftarrow \mathcal{D}}[|R' \cap S'| = z] &= \sum_{Z \subseteq [n'], |Z| = z} \Pr_{R',S' \leftarrow \mathcal{D}}\left[(R'_Z = S'_Z) \wedge \left(R'_{\overline{Z}} \cap S'_{\overline{Z}}) = \emptyset\right)\right] \\ &\leq \sum_{Z \subseteq [n'], |Z| = z} \Pr_{R',S' \leftarrow \mathcal{D}}[R'_Z = S'_Z] \leq \sum_{Z \subseteq [n'], |Z| = z} \left(\frac{1}{n^{1.5}}\right)^z \end{aligned}$$

The second inequality holds since each element in the stream has min-entropy at least $\lg(n^{1.5})$. Therefore, we have $\Pr_{R',S' \leftarrow \mathcal{D}}[|R' \cap S'| = z] \leq \binom{n/3}{z} \cdot \left(\frac{1}{n^{1.5}}\right)^z \leq \left(\frac{1}{n^{0.5}}\right)^z$.