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

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

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

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 investigate this folklore to quantify the privacy of the min-hash sketch.

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. This immediately yields a privacy-preserving sublinear-communication semi-honest 2-PC protocol based on FHE where the hash function is evaluated homomorphically.

To improve the efficiency of this protocol, we next consider an implementation in the random oracle model, i.e., in the presence of a *public* ideal hash function. Here, the protocol participants jointly sample public prefixes for domain separation of the random oracle, and locally evaluate the resulting hash functions on their input sets. 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 then the output of the min-hash functionality achieves DDP, again without any added noise. This yields a more efficient semi-honest two-party protocol in the random oracle model, where parties first locally hash their input sets and then perform a 2PC for comparison.

By proving that our protocols satisfy DP and DDP respectively, our results formally confirm and qualify the folklore belief that min-hash based protocols protect the privacy of their inputs.

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 [11,43]. The Jaccard index [39] is a similarity measure between two sets A and B, denoted $J(A, B)$, which is defined as the fraction of the elements in the intersection of A and B divided by the number of elements in the union of them. 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 [11,12], community identification [63], DNA matching [16], and machine learning [65,40].

However, computing the Jaccard index exactly, especially when the input sets are large, can be costly. The min-hash sketch allows communication-efficient approximation. The basic idea behind the min-hash sketch is to apply a random permutation π to both sets A and B and then to see whether the last item in both sets (under this permutation) is the same. Since this permutation is applied over all elements in $A \cup B$, it is easy to see that the last item will be the same exactly when the last item over $A \cup B$ is in both A and B. Specifically,

$$\Pr[\text{last item in } \pi(A) = \text{ last item in } \pi(B)] = \frac{|A \cap B|}{|A \cup B|} = J(A, B)$$

Thus, to get an unbiased approximation of the Jaccard index, it suffices to repeat this procedure with sufficiently many random permutations.

Motivation. Due to its simplicity and efficiency, the min-hash sketch has become a very popular tool to approximate the Jaccard index. Moreover, since the min-hash sketch only needs to compare the last items in the permuted sets, 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. [16,10,54,28]) 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 implication of revealing the final result. The second line of work (e.g. [66,67,1]) considers how to make the min-hash itself more privacy-preserving. Specifically, these works consider adding noise to the output of the min-hash algorithm to guarantee differential privacy (DP) of individual items in the input sets.

These works serve as the starting point for our paper. We aim to design protocols for two-party secure estimation of the Jaccard index that additionally preserve the privacy of individual items inside the input sets. But, unlike prior work, we make a very important restriction. To minimize the error of the approximation and to guarantee reproducibility of the results, we do not allow our protocol to add any additional noise to the output of the min-hash protocol. This goal forces us to ask the following critical question:

> *Does the min-hash algorithm itself provide privacy guarantees for its inputs?*

Differential privacy of the min-hash sketch. Somewhat surprisingly, we show that the answer to this question is yes. Specifically, we show that under certain application scenarios and restrictions on the input sets, the error of the min-hash approximation of the Jaccard index is actually sufficient to achieve differential privacy. Essentially, the error of the sketch acts as noise to protect the privacy of the inputs. Similar observations that sketching algorithms inherently preserve privacy have previously been shown for the Johnson-Lindenstrauss sketch [9], the LogLog sketch [15,59], and other sketches [64].

To get an initial understanding of the underlying privacy of the min-hash protocol, we first consider a setting with a set of private permutations chosen by the functionality and unknown to the parties. The functionality then uses these permutations to find the minimum item in the two sets and output the total number of times the minimums match. 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 1 (Informal). *If the size of the intersection is a constant fraction of the size of A and B, then the output of this min-hash protocol 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 permutations, and cannot learn the evaluation of the permutations on their own inputs. This, of course, poses a particular challenge if one wants to deploy this protocol without relying on a trusted party or functionality. However, despite this challenge, we note that using fully-homomorphic encryption [32] to evaluate the permutations and compare the minimums, it is possible to build a protocol with sublinear communication. Since it securely realizes the functionality (with semi-honest security), the protocol would achieve computational DP [7,23,50,56]. Additionally, since the error is unbiased and from an easily sampleable distribution, this protocol is also a secure approximation [29] of the Jaccard index in that it leaks no information about the input sets beyond the Jaccard index.

Distributional DP in the public hash setting. However, the need to keep the permutations private in Theorem 1 forces our initial protocol to evaluate the permutations inside of FHE for each input, resulting in an inefficient protocol in practice.

Therefore, for our main results, we consider what happens *if we do not keep the permutations private from the parties.* The major benefit of revealing the permutations is that the parties can then apply them to their own sets locally, and only perform secure computation to compare the minimum elements and count the number of matches. Thus, allowing for much simpler protocols with *sublinear communication* and only $\tilde{O}(k)$ secure gate evaluations where k is the number of permutations which is far smaller than the input size.

Unfortunately, there is a problem with the above approach. It is easy to see that, if we reveal the permutations, the min-hash protocol is not in fact differentially private, as defined above. The problem stems from the fact that 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 permutations are known, then the adversary can reconstruct the min-hash exactly both for the case when x^* is in the set and when it is not, and then 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.

The above gives a counter-example against the differential privacy of min-hash in the public permutation setting. However, does it really give rise to a realistic privacy breach?

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, the adversary would not know the entire input of the honest party. Indeed, parties would only want to perform a secure computation on their private inputs in the first place if they did not already know what the answer would be. In the context of min-hash, this means that the adversary does not already know a good portion of the honest party's input set. More precisely, this motivates the assumption 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) [4] 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 2 (Informal). *If each non-intersecting item has sufficiently high min-entropy, revealing the hash functions*[4] *together with the min-hash counts 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 4.2) 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,

[4] We use cryptographic hash functions to instantiate the permutations in the random oracle model.

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 super-logarithmic 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 the min-hash protocol still achieves DDP, by proving a new strong chain rule for min-entropy (see Section 7).

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.
- Arbitrary leakage $L = L(R)$ is computed on R, we require 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.

Theorem 3 (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 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 [26]). To deal with such situations, Skórski [57] 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, if we want to leak $\{\ell_i\}_{i \in W}$ *before* computing the spoiling bits, we are not able to do so, since we don't know W yet! We could leak from all the blocks $(R_1, \ldots, R_n)$, but then it may destroy the properties 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 4.

A note on composition. One known weakness of the DDP definition is the lack of a general composition theorem [4]. However, for the specific setting of our min-hash protocols we can leverage the small output of min-hash to argue composition properties of our protocols. 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.

Adversary model. All of the protocols described above are secure in the random oracle model against a semi-honest adversary corrupting one of the parties. Achieving malicious security, especially for the DDP protocol, is an interesting open question.

Comparison to other approaches for Jaccard index estimation. We note that an alternative approach to get a differentially-private estimate of the Jaccard index is via mergeable cardinality estimation sketches (e.g. [37]) to compute (an approximation of) the set intersection cardinality and use this via the inclusion-exclusion principle to compute the Jaccard index. However, all such cardinality sketches we are aware of rely on adding independent noise to achieve DP resulting in an additional source of error. We give a detailed comparison of error from our protocol vs. the best known cardinality estimator [37] in Section 9.

Our results, on the other hand, show that in many cases, the min-hash sketch is already sufficient to achieve differential privacy of the inputs. This result allows us to build low-error and sublinear communication protocols for approximating the Jaccard index without adding any additional noise to achieve privacy. Thus, in addition to preserving privacy, results produced by our protocols are reproducible, generating the same output if run with the same inputs and the same hash functions.

Finally, this leads to the following encouraging observation. The min-hash has been in use for a long time, and in that time people have not really considered whether the results they were computing and storing were privacy preserving. As we now show that the min-hash sketch itself is DP, this gives evidence that prior computation may already preserve privacy of the individual inputs.

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 [22,24].

A large body of works have developed differentially private algorithms [25] for a variety of computations. Most of these works focused on the standard setting with a trusted curator who has access to all users' data and aims to respond in a differentially private manner.

Differential privacy has also been considered in a multi-party setting. In this setting two common approaches for differential privacy are local-DP, where parties add noise to their own inputs prior to performing the computation (e.g., [27,41]) or secure computation emulating the trusted curator (e.g., [8])

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. Beimel et al. [8] first proposed such optimization for the problem of secure summation. He et al. [36] and Groce et al. [34] applied the differential privacy relaxation to improve efficiency of set-intersection protocols. Mazloom and Gordon [46], and Mazloom et al. [47] consider graph-parallel computations and design more efficient solutions with differential private leakages. Chan et al. [13] consider classic tasks like sorting, merging, and range-query data structures with differential privacy relaxation. Gordon et al. [33] 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 [9,15,59,64,37,19,44,53,49,48,6,5,38,69] 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 [61,60,51,42,52]. Recently, Hehir et al. [37] proposed a private mergeable sketch that can be used to estimate the size of the intersection and union of sets.

We note that our work achieves reproducibility, since it adds no noise. There exists earlier work trying to achieve reproducibility by using sticky noise (i.e., adding noise, but deriving the noise from the data) [18,30]; however, simply following this approach to achieve reproducibility may cause a system to be susceptible to attacks [31].

Secure Approximation. Secure approximation studies what functions can be securely approximated without revealing anything beyond the true output [29,35]. While this notion is quite different from that of differentially-private approximation that we consider here, we note that our first (FHE-based) protocol additionally achieves this.

3 Preliminaries

Notations. 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. 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.

Hash functions in 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 $\mathsf{negl}(\kappa)$.

Differential privacy. We first give the definition of the traditional (ϵ, δ)-differential privacy.

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. $x \simeq_X x' :\ \mathcal{M}(x) \overset{c}{\approx}_{\epsilon,\delta} \mathcal{M}(x')$.

Distributional differential privacy (DDP). We adapt the original definition [4] for our purpose to consider a two-party protocol that takes sets as input more explicitly. Specifically, we consider the following computational indistinguishability variant for our DDP definition.

Definition 4 (View of a party in a two-party protocol). *Given a two-party protocol Π with parties P_1 and P_2, let $\mathsf{view}_{P_1}^{\Pi}(A, B)$ denote the view of P_1 for the execution of protocol Π with A and B being the input of P_1 and P_2 respectively. In particular, $\mathsf{view}_{P_1}^{\Pi}(A, B)$ consists of the following:*

- *The input A of P_1, the randomness that P_1 uses, the messages that P_1 receives from P_2, and the output of the protocol.*
- *If the protocol 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.*

The view of P_2 is defined similarly.

We now define computational DDP for a two-party protocol against an adversary that can corrupt one of them in an honest-but-curious manner.

Definition 5 (Computational DDP of a two-party protocol). *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 protocol Π 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 $x^* \in \mathcal{U}$, the following holds:*

$$\left(\mathsf{view}_{P_1}^{\Pi}(A, B), Z\right) \overset{c}{\approx}_{\epsilon,\delta} \left(\mathsf{view}_{P_1}^{\Pi}(A, B_{+x^*}), Z\right).$$

In the above, (A, B) and Z are sampled from Δ, and each party may use additional randomness. DDP against an adversary corrupting P_2 is defined symmetrically.

Finally, the following lemma upperbounds the tail of a Binomial distribution.

Lemma 1 ([21]). *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 Our Protocols

4.1 Min-Hash Protocol in the Private Hash Setting

In Figure 1, we describe the standard min-hash protocol in the private hash setting ($\mathcal{F}_{\mathsf{privH}}$). In particular, after choosing k random hash functions, the mechanism computes the number of iterations in which the min-hash of A matches the min-hash of B.

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

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 k hash functions $h_1, h_2, \ldots, h_k$.
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, $u_j^A = \min_i\{y_{i,j}^A\}$
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 c to P_1 and P_2.

Fig. 1: Min-Hash in the Private Hash Setting

Theorem 1. *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 $\delta \in \mathsf{negl}(\kappa)$.*

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 $(u_1^A, \ldots, u_k^A)$ (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.

4.2 Two-party Public Min-Hash Protocol with DDP

In Figure 2, we describe a two-party protocol $\pi^{\mathcal{O}}_{\mathsf{PH}}$ in the random oracle (RO) model achieving DDP. The protocol is essentially the same as $\mathcal{F}_{\mathsf{privH}}$ except for the following differences:

- Prefixes $\mathsf{pre}_1, \ldots, \mathsf{pre}_k$ are jointly sampled by the parties.
- The hash functions h_i for $i \in [k]$ are defined as $h_i(\cdot) := \mathcal{O}(\mathsf{pre}_i || \cdot)$.
- A generic secure two party computation [68,45] is used to hide the information of u_i^As and u_i^Bs in the equality check.
- The parties' output consists of the min-hash output, as well as the pre-fixes $\mathsf{pre}_1, \ldots, \mathsf{pre}_k$. This additional information greatly impacts the analysis.

The main benefit of the protocol is that the hash computations can be computed locally in the clear without FHE.

Protocol $\pi^{\mathcal{O}}_{\mathsf{PH}}$

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. Both parties jointly sample k prefixes $\mathsf{pre}_1, \ldots, \mathsf{pre}_k$. These are 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. Party 1 takes its input A from which it computes the min-hash vector $(u_1^A, u_2^A, \ldots, u_k^A)$ exactly as described in $\mathcal{F}_{\mathsf{privH}}$. Likewise, Party 2 computes $(u_1^B, u_2^B, \ldots, u_k^B)$ similarly using its input B.
3. Taking $(u_1^A, u_2^A, \ldots, u_k^A)$ and $(u_1^B, u_2^B, \ldots, u_k^B)$ as input, both parties execute a generic secure-two party computation protocol for computing $c := \sum_{j=1}^{k} \mathsf{Eq}(u_j^A, u_j^B)$.

Output: Return c, $\mathsf{pre}_1, \ldots, \mathsf{pre}_k$, to P_1 and P_2.

Fig. 2: The Protocol for Min Hash in the RO Model

Distributional Differential Privacy. In Figure 3, we 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).

We assume that each of the non-intersecting elements has high min-entropy. WLOG, consider an adversary corrupting P_1. We can safely ignore the protocol messages in Step 3 in our analysis, thanks to the security guarantees of secure two-party computation. Therefore, the view of the adversary will be

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

Distribution Family Δ_{PH}

This family of distributions is parameterized as follows:

> The universe $\mathcal{U}$ for the elements in sets A and B; the cardinalities n_A, n_B, n_I of sets A, B, $I = A \cap B$ respectively; distribution $\mathcal{D}_{A,I,B'}$ for sampling (A, I) and $B' = B \setminus A$.

Choose (A, I, B') according to $\mathcal{D}_{A,I,B'}$, which should satisfy the following requirements:

– $|A| = n_A, |I| = n_I, |B'| = n_B - n_I, I \subset A, A \cap B' = \emptyset$.

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.

Fig. 3: The family of distributions that we consider in our min-hash protocol

As with the case of $\mathcal{F}_{\mathsf{privH}}$, since the hash functions are chosen at random, we can apply the same analysis and show that 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.

Hashes of non-intersecting items working as a noise. First, since the hash functions are public, iterations are no longer independent of each other as needed by the analysis in Section 4.1. 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.

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. Let $n_R = n_B - n_I$, and the above discussion can be formalized into the following definition.

Definition 6 (θ-good iteration). *We define a predicate* $\mathsf{good}_\theta(h_j, A, I, n_B)$ *to be true if and only if the following holds:*

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

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 v_j^R may reduce the final count with a decent chance, and it may not with a decent chance too. 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 $k_g = \Theta(k)$.

Since the hash functions are now public, and therefore $\min h_j(I)$ is 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, we can still show that this distribution works as a good noise to hide the private data. In particular, we use the techniques of [14] to upper-bound the privacy guarantee as those from a binomial mechanism (with rather worse parameters compared to $\mathcal{F}_{\mathsf{privH}}$).

Theorem 2. *For every constant* $\epsilon > 0$*, consider protocol* π_{PH} *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, protocol* π_{PH} *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.*

4.3 DDP for Input Sets Chosen from a Poly-sized Universe

Let $n_R = n_B - n_I$. In this section, we show that π_{PH} 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 $L \leq n_R(\lg \ell - 3 \lg n_R - 2)$. Assume without loss of generality that the adversary corrupts P_1.

Noise: fixed secret set over the choice of hashes. Consider a fixed secret set $R' \subset R = B \setminus I$ (we will see how to choose R' later). Let $n = |R|$ and $n' = |R'|$ and consider the probability $E_r^{R'}$ over choice of the hash functions that R' contributes noise of $-r$ to the total count c in the min-hash protocol over a bundle of k_b good iterations (we will see how to choose k_b later). We can show that the distribution over r has the following property: There exist $a, b \in \{0, \ldots, k_b\}$ such that:

- The probability of obtaining $r \notin [a, b]$ is negligible.
- For every $\epsilon \in [0, 1]$ and sufficiently large security parameter κ, for every $r \in [a, b]$, we have $e^{-\epsilon} \leq \frac{\Pr[r]}{\Pr[r+s]} \leq e^{\epsilon}$ where $s = \lg \lg \kappa$ is a small value corresponding to the sensitivity.

The above indicates that the distribution over r, when the set R' is fixed and the probability is taken over choice of hash functions, is amenable for use as a noise distribution in a differential privacy context. It turns out $E_r^{R'}$ depends on only the size of R' and For every r, we let $E_r^{n'}$ denote its probability under the distribution described above.

Noise: fixed hash over the choice of the secret set. Our main technical challenge is to show that the properties needed for differential privacy hold *even when we switch quantifiers.* Specifically, given a distribution $\mathcal{D}$ over sets R' and a *fixed* hash function (modeled as a random oracle), let d_r be the probability over R' chosen from $\mathcal{D}$ that R' contributes $-r$ to the total count in the min-hash protocol. We would like to show that for any fixed hash functions, for all $r \in [a, b]$, it holds $e^{-\epsilon} E_r^{n'} \leq d_r \leq e^{\epsilon} E_r^{n'}$. Then, we can meaningfully upper-bound the distance between d_r and d_{r+s} using the property of $E_r^{n'}$ described above.

The aforementioned universal quantifier for the hash functions can be slightly relaxed by requiring the property to hold over the choice of the hash functions with all but small probability (we will later describe how to compensate the relaxation by the bundling technique). In particular, let D_r be the random variable corresponding to probability d_r over the choice of hash functions. We would like to show that with all but small probability over the choice of the hash functions, for all $r \in [a, b]$, it holds $e^{-\epsilon} E_r^{n'} \leq D_r \leq e^{\epsilon} E_r^{n'}$.

Geometric collision Property. To show the above, our high-level strategy is to use Chebyshev's inequality, which gives good concentration bounds on D_r when $\mathsf{Var}[D_r]$ is bounded. Thus, our next goal is to upperbound $\mathsf{Var}[D_r]$. 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 $\mathsf{Var}[D_r]$ can be bounded for any distribution over sets R' that has this property.

Geometric collision property in the face of leakage. 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 [20, Lemma 2.2]; if we want 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 [58]. 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 [58] to produce spoiling leakage with the following properties:

- It can be computed element-by-element, starting from the last element of the distribution.
- Conditioned on the spoiling leakage, we can simulate the adversary's entire view in the protocol (by leaking extra information along with the spoiled bits), other than the outcome of the random variable r over choice of R, where R is drawn from the probability distribution that conditions on the spoiling leakage.
- Conditioned on the spoiling leakage, each element R_i either has max-entropy at most $\sim 1.5 \cdot \lg(n)$ or min-entropy at least $\sim 1.5 \cdot \lg(n)$.

We further show that, due to the high min-entropy of the entire set R, there must be a constant fraction of blocks R' with min-entropy at least $1.5 \cdot \lg(n)$. To reflect the fact that only a constant fraction of elements are guaranteed to have high min-entropy, the parameter n' is set to $n/3$. We now consider the marginal distribution over R' (the elements of R with high min-entropy) and show that the geometric collision property holds with respect to this distribution.

Bundling iterations towards DDP with negligible δ. We are not quite done yet. By applying Chebyshev 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, and argue that w.h.p. Chebyshev succeeds for at least one bundle. Note that the hash functions 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 superconstant u. For this, we set the parameter $k_b = k_g/u$, where k_g is the number of good iterations. The fact that the probability distribution of r over choice of $\mathcal{R}$ is good for at least one bundle (with all but negligible probability) is sufficient to complete the proof of distributional differential privacy.

Theorem 3. *For security parameter κ, every constant $\epsilon > 0$, and every constant $\gamma \in (0, 1)$, consider protocol π_{PH} in Figure 2 in the random oracle model with $k = k(\epsilon, \kappa)$, where $k \in \Omega(\kappa \cdot \lg \lg \kappa)$. Let $R = B \backslash I$ 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 π_{PH} 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.*

5 DP of $\mathcal{F}_{\mathsf{privH}}$: Proof of Theorem 1

Setting the scene. WLOG, we consider two neighboring inputs

$$(A, B) \text{ and } (A, B_{+x^*}).$$

Differential privacy for the case in which x^* is added into A can be shown symmetrically. WLOG, let $B = (x_1^B, \ldots, x_{n_B}^B)$ and $B_{+x^*} = (x_1^B, \ldots, x_{n_B}^B, x^*)$.

Sensitivity. 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. Since we model each hash function h_j as a random oracle, $(y_{1,j}^B, \ldots, y_{n_B+1,j}^B)$ will be uniformly distributed. We observe the following:

> *Consider how the min-hash u_j^B is computed. The value x^* from B_{+x^*} can affect the min-hash u_j^B (and thereby the final count c), only if $y_{n_B+1,j}^B$ is smaller than $(y_{1,j}^B, \ldots, y_{n_B,j}^B)$.*

The probability that $y_{n_B+1,j}^B$ will be the minimum is $1/(n_B+1)$ by a symmetry argument. Note the final output is computed as the sum of k of these trials. Let

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

Therefore, we consider a binomial distribution as follows

$$|S_{x^*}| \sim \mathrm{B}(k, 1/(n_B + 1)),$$

which represents how many iterations j cause x^* to be the min-hash u_j^B. In other words, $|S_{x^*}|$ captures the sensitivity of min-hash.

The following lemma upper bounds this sensitivity.

Lemma 2. *For any $\{x_i^B\}_{i \in [n_B]}$ and any $x \in \mathcal{U}$, we have $\Pr_{h_1, \ldots, h_k}[|S_x| \geq s] \leq \left(\frac{e \cdot k}{s \cdot (n_B+1)}\right)^s$.*

Proof. We have $\Pr[|S_x| \geq s] \leq \binom{k}{s} \cdot \left(\frac{1}{n_B+1}\right)^s \leq \left(\frac{e \cdot k}{s}\right)^s \cdot \left(\frac{1}{n_B+1}\right)^s$. The first inequality is from Lemma 1. □

Remark. From the above lemma, for $k \in \Theta(\kappa)$, $n_B \in \Omega(\kappa^2)$, we have

$$\Pr[|S_{x^*}| \geq \lg\lg\kappa] \leq \mathsf{negl}(\kappa).$$

This implies that given the parameters above, with overwhelming probability, there are at most $\lg\lg\kappa$ iterations where x^* changes the min-hash value for B.

Binomial distribution hides the small sensitivity. Let $s = \lg\lg\kappa$ and let $p = J(A, B)$. Let $K_{x^*} = [k] \setminus S_{x^*}$. For the iterations in K_{x^*}, the min-hash matches for both (A, B) and (A, B_{+x^*}) will be identically distributed.

Let $c_{K_{x^*}}$ denote the match count in iterations K_{x^*}. Note that since we model each hash function as a random oracle, 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* $\mathrm{B}(n, p)$*, where* $n \in \Omega(\kappa)$ *and* $p \in (0, 1)$ *is a constant independent of* κ*. Then, for any constant* ϵ *and* $s \leq \lg\lg\kappa$*, there are* $a, b \in [n]$ *with* $a < np < b$ *such that*

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

6 DDP of π_{PH}: Proof of Theorem 2

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. We first define the following ideal functionality $\mathcal{F}_{\mathsf{PH}}$ in the random oracle (RO) model.

Functionality $\mathcal{F}^{\mathcal{O}}_{\mathsf{PH}}$

$\mathcal{F}^{\mathcal{O}}_{\mathsf{PH}}$ works exactly the same as $\mathcal{F}_{\mathsf{privH}}$ except that (1): It internally samples random prefixes $\mathsf{pre}_1 \ldots, \mathsf{pre}_k$ and in the i-th iteration, hashes the parties' input vectors using $h_i(\cdot) = \mathcal{O}(\mathsf{pre}_i||\cdot)$. (2) It outputs c, $\mathsf{pre}_1 \ldots, \mathsf{pre}_k$. Relative to public random oracle $\mathcal{O}$, this allows P_1 and P_2 to evaluate $h_1, h_2, \ldots, h_k$.

Fig. 4: The Ideal Functionality for Public Min Hash

We also define a slightly different ideal functionality $\mathcal{F}^{(1)}_{\mathsf{PH}}$ as follows:

- Let $\mathcal{F}^{(1)}_{\mathsf{PH}}$ be the same as $\mathcal{F}_{\mathsf{PH}}$ except that for each $x^B_i \in B \setminus A$, each element in $\{y^B_{i,j}\}_j$ 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

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

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.

Thanks to the standard technique of the generic two-party secure computation [45], the following holds:

$$\mathsf{view}_{P_1}^{\pi_{\mathsf{PH}}}(A,B), I) \overset{c}{\approx} (\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{PH}}}(A,B), I), \quad \mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{PH}}}(A,B_{+x^*}), I) \overset{c}{\approx} (\mathsf{view}_{P_1}^{\pi_{\mathsf{PH}}}(A,B_{+x^*}), I).$$

Now we show $(\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{PH}}}(A,B), I) \overset{c}{\approx} (\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{PH}}^{(1)}}(A,B), I)$. 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 $\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{PH}}}(A,B_{+x^*}), I) \overset{c}{\approx} (\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{PH}}^{(1)}}(A,B_{+x^*}), I)$.

6.1 DDP of $\mathcal{F}_{\mathsf{PH}}^{(1)}$

From the above discussion, we are only left to show

$$(\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{PH}}^{(1)}}(A,B), I) \approx_{\epsilon,\delta} (\mathsf{view}_{P_1}^{\mathcal{F}_{\mathsf{PH}}^{(1)}}(A,B_{+x^*}), I).$$

In other words, we need to 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 $\mathcal{F}_{\mathsf{PH}}^{(1)}(A,B)$ and c_{+x^*} is the final count from $\mathcal{F}_{\mathsf{PH}}^{(1)}(A,B_{+x^*})$.

In Section 5, we studied the sensitivity of the final count when adding an element x^* to B. 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.

The following lemma shows that in the random oracle model, 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.

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

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

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 $|S_{x^*}| \le \lg\lg\kappa$).

Now, 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*

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

Fixing randomness. Recall we aim to show that

$$(A, I, h_1, \ldots, h_k, c) \approx_{\epsilon,\delta} (A, I, h_1, \ldots, h_k, c_{+x^*}). \tag{1}$$

To achieve this goal, we divide the output count into two parts:

- The part corresponding to K_θ, which contains all the θ-good iterations such that x^* does not hash to the minimum across B_{+x^*}.
- The part contributed by all the remaining iterations.

We will first argue that the difference of the second parts for the neighboring inputs B and B_{+x^*} is upper-bounded by a small value, following our discussion on sensitivity in Section 5. Then we will treat the first part as the noise distribution and derive the privacy guarantee it provides to hide the difference from the second part.

For a set $W \subset [k]$, define its complement as $\overline{W} = [k] \setminus W$. In addition, we introduce the following notations.

- For sets V, W, let $Y_{V,W} \overset{def}{=} \{y_{i,j}^B : i \in V, j \in W\}$.
- For a set W, define $c_W \overset{def}{=} \sum_{j\in W} \mathsf{Eq}(u_j^A, u_j^B)$ and

Now, fix A, I, x^* and $h_1, \ldots, h_k$. Recall that we are considering $\mathcal{F}_{\mathsf{PH}}^{(1)}$ that satisfy the following condition:

- For each $x_i^B \in B \setminus A$, each element in $\{y_{i,j}^B\}_j$ is chosen uniformly at random from $[0, 1]$.

Therefore, even after $(A, I, x^*, h_1, \ldots, h_k)$ is fixed, it holds that $Y_{R,[k]}$ are still uniformly distributed. Moreover, we additionally fix $Y_{R,\overline{G}_\theta}$, which implies that only Y_{R,G_θ} is uniformly random.

Note that fixing $A, I, x^*, h_1, \ldots, h_k$ also allows us to determine which iterations are θ-good (i.e., G_θ).

Small sensitivity. Given all this, we first show that there for $s = \lg\lg\kappa$,

$$\Pr_{Y_{R,G_\theta}}\left[\left|c_{\overline{K}_\theta} - c_{\overline{K}_\theta}^{+x^*}\right| > s\right] \le \mathsf{negl}(\kappa).$$

To show inequality, noting that $\overline{K}_\theta = \overline{G}_\theta \cup S_{x^*}$, consider two cases:

– For iteration $j \in \overline{G}_\theta \setminus S_{x^*}$, we have $u_j^B = u_j^{B_{+x^*}}$. This is because $j \notin S_{x^*}$ implies that x^* doesn't lead to min-hash. Therefore, such iterations don't contribute to the difference of c and c^{+x^*}.
– For the rest iterations $j \in S_{x^*}$, we have $|S_{x^*}| \leq s$ with overwhelming probability due to Lemma 2, which shows the above equation.

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

$$\Pr_{Y_{R,G_\theta}}[c_{\overline{K}_\theta} + c_{K_\theta} = q] \text{ and } \Pr_{Y_{R,G_\theta}}[c^{+x^*}_{\overline{K}_\theta} + c^{+x^*}_{K_\theta} = q].$$

Note that we have $c_{K_\theta} = c^{+x^*}_{K_\theta}$ because $j \in K_\theta$ implies $j \notin S_{x^*}$. Therefore, based on the sensitivity argument shown above, we only need to analyze the distribution of c_{K_θ} and compare the following two probabilities:

$$\Pr_{Y_{R,G_\theta}}[c_{K_\theta} = q] \text{ and } \Pr_{Y_{R,G_\theta}}[c_{K_\theta} + s = q].$$

Distribution of c_{K_θ}. We abuse notation and denote $c_j = c_{\{j\}}$. Then, we have $c_{K_\theta} = \sum_{j \in K_\theta} c_j$. Let $Y_{R,j} = \{y^B_{i,j} : i \in (n_I, n_B]\}$. Note that since we have $j \in K_\theta$, a θ-good event takes place in iteration j, i.e., $\min h_j(A) = \min h_j(I)$. This implies the following:

If $\min Y_{R,j} \geq \min h_j(I)$, *we have* $c_j = 1$*; otherwise we have* $c_j = 0$.

Let $\gamma_j = 1 - \min h_j(I)$. Then, the probability that $c_j = 1$ is $(\gamma_j)^{n_R}$, since every number in $Y_{R,j}$ must be be greater than or equal to $\min h_j(I)$. Let $\eta_{-\theta} = 1/2 - \theta$ and $\eta_{+\theta} = 1/2 + \theta$. Note that since $j \in G_\theta$, the following holds according to Definition 6:

$$(\gamma_j)^{n_R} \in [\eta_{-\theta}, \eta_{+\theta}].$$

Therefore, letting $p_j = (\gamma_j)^{n_R}$, we have $c_j \sim \textsc{Ber}(p_j)$, where $\textsc{Ber}$ denotes the Bernoulli distribution. Also, recall that every number in $Y_{R,j}$ is chosen uniformly at random from $[0, 1]$ and therefore these Bernoulli distributions are independent from each other. Therefore, we can apply Lemma 6 below to conclude that $C_{K_\theta} \approx_{\epsilon,\delta} C_{K_\theta} + s$.

DP of Additive Poisson Binomial distribution. 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.

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* $s \leq \lg\lg\kappa$*, 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* $\ell \notin [a + s, b]$, $\Pr[\mathsf{PB}(n, p_J) = \ell] = \mathsf{negl}(\kappa)$ *and* $\Pr[\mathsf{PB}(n, p_J) + s = \ell] = \mathsf{negl}(\kappa)$.

7 Strong Chain rule

Our proof of Theorem 3 requires a special case of the following theorem, corresponding to a strong chain rule for specific leakage functions $\ell_1(\cdot), \ldots, \ell_n(\cdot)$. By formalizing the properties needed from $\ell_1(\cdot), \ldots, \ell_n(\cdot)$ for the proof of the special case to go through, we are able to arrive at the generalization presented in this section. We state the generalized version of the theorem since we believe it may find future applications in leakage-resilient cryptography.

Recall that we consider a block-by-block random variable $R = (R_1, \ldots, R_n)$, and (potentially randomized) leakage functions $\ell_1(\cdot), \ldots, \ell_n(\cdot)$ with randomness $\rho_1, \ldots, \rho_n$. You can think of the blocks as coming in a streaming fashion in order of $R_1, R_2, \ldots, R_n$.

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 as well as all the outputs of $(\ell_{i+1}, \ell_{i+2}, \ldots, \ell_n)$ and that the total number of valid sequences of leakages from $\ell_1(\cdot), \ldots, \ell_n(\cdot)$ is sufficiently small (see Property 1 in Theorem 4).

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

The properties of $(R_1, \ldots, R_n)$ and $(\rho_1, \ldots, \rho_n)$ 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 4).

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 $\ell_1(\cdot), \ldots, \ell_n(\cdot)$ 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 4)

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 4 (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* $\Pr_R[f(R) = \perp] \leq \epsilon n$.
3. $|Im(f)| \leq B \cdot (2(\lg(\ell) + \lg(1/\epsilon))/\delta)^n$.
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.,*

$$\forall y \in Im(f) \setminus \{\perp\}, \forall r \text{ s.t. } f(r) = y, \forall i \in V :$$
$$\Pr\left[R_i = r_i \;\middle|\; R_{<i} = r_{<i},\ y\right] \leq \frac{2^\delta}{c}.$$

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.,*

$$\forall y \in Im(f) \setminus \{\perp\}, \forall r \text{ s.t. } f(r) = y, \forall i \in W :$$
$$|\{r_i : \Pr[R_i = r_i | R_{<i} = r_{<i}, y]] \geq 0\}| \leq 2^\delta \cdot c.$$

8. $\{\rho_i\}_{i \in V}$ *are distributed independently and uniformly at random conditioned on* $f(R)$.

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 the setting of c and the size of 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 $|V| \geq n/3$. We show that our min-entropy assumption on R implies that this parameter setting is achievable in Section F.3.

8 Highlights of Proof of Theorem 3

Due to the lack of space, we highlight only the important parts of the proof of Theorem 3. The full proof can be found in Appendix D.

Remember that the adversary holds set A of size n_A. The honest party holds set B of size n_B. The intersection $I := A \cap B$ has size n_I. The secret set held by the honest party $R := B \setminus I$ has size n_R.

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 π_{PH} 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* B. A distribution of B is equivalent to a distribution of how to choose n_B nodes in $\mathcal{X}$.
2. Use H_1 to determine G_θ and set $\mathcal{Y} = G_\theta$. In other words, $\mathcal{Y}$ contains all good iterations that could potentially increase the final count.
3. Let $p_j = \min h_j(I)$. Use H_2 to determine the set of edges:
$$\mathcal{E} = \{(i, j) : (i, j) \in \mathcal{X} \times \mathcal{Y} \text{ and } h_j(x_i) < p_j\}.$$
In other words, existence of an edge (i, j) means that if node i belongs to input B, iteration j will not contribute to the final count.
4. Output the resulting bipartite graph $(\mathcal{X}, \mathcal{Y}, \mathcal{E})$.

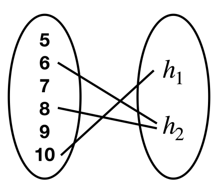

Fig. 5: Min-hash graph

	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 5 shows the constructed min-hash graph. In particular, we have $\mathcal{X} = \{5, 6, \ldots, 10\}$. We have $\mathcal{Y} = \{h_1, h_2\}$; 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 of Secret Items and Good Iterations

We use the min-hash graph and analyze DDP of π_{PH} for subsets of secret items and good iterations. We first *fix items* and consider a noise distribution *over the choice of hash functions*. Extending this, in the next section, we will consider the noise over a distribution of items.

Edges in a min-hash graph. In the random oracle model, each answer in H_2 is chosen uniformly at random. Therefore, given H_1, the probability (over the choice of H_2) that an edge (i, j) forms is exactly equal to p_j. Moreover, the probability that (i, j) forms is independent of the probability that any other edge in the graph forms.

Distribution of min-hash graphs over the choice of H_2. Fix the hash answers H_1 to fix $\mathcal{X}$ and $\mathcal{Y}$. Fix any positive integer $\hat{n} \le n_R$ and the set $T \subseteq \mathcal{Y}$ of iterations with $|T| = \hat{k} \le k_g$. For a fixed set $\hat{R} \subset \mathcal{X}$ of $\hat{n}$ nodes on the left, let $\mathcal{G}_Y^{\hat{R}}$ be the set of all min-hash graphs that cover *exactly* the set $Y \subseteq T$. More formally, given a set of edges $\mathcal{E}$, define the set of destination nodes as $\mathsf{dest}(\mathcal{E}) = \{j : (i,j) \in \mathcal{E}\}$ respectively. Define the set of all possible min-hash subgraphs in which the set of destination nodes is exactly Y (from $\hat{R}$).

$$\mathcal{G}_Y^{\hat{R}} = \{(\hat{R}, Y, \mathcal{E}') : \mathcal{E}' \in \hat{R} \times Y, \mathsf{dest}(\mathcal{E}') = Y\}.$$

In this section, we are interested in *the probability over the choice of the hash functions that the final count is reduced by exactly r due to the elements of a fixed set $\hat{R}$ over the $\hat{k}$ iterations in T.* In particular, we define

$$E_{T,r}^{\hat{R}} = \sum_{Y \subseteq T, |Y| = r} \sum_{G \in \mathcal{G}_Y^{\hat{R}}} p(G),$$

where $p(G)$ is defined as

$$p(G) = \Pr_{H_2}[G_0 \leftarrow \mathbf{MinhashG}_{H_1}(A, I, x^*, H_2); G \text{ is a subgraph of } G_0].$$

As explained above, in the random oracle model, the probability depends only on the size of the sets $\hat{R}$ and T (i.e., not the actual identity of the set). Therefore, we will often use the notation $E_{\hat{k},r}^{\hat{n}} = E_{T,r}^{\hat{R}}$ when $|\hat{R}| = \hat{n}$ and $|T| = \hat{k}$.

Observe that $E_{\hat{k},r}^{\hat{n}}$ is another way of representing an Additive Poisson Binomial distribution. That is,

$$E_{\hat{k},r}^{\hat{n}} = \Pr[\mathsf{PB}(\hat{k}, p_J) = r].$$

Therefore, based on Lemma 6, we have the following:

Corollary 1. *Fix H_1. Let $s = \lg\lg\kappa$. For any constant ϵ, and any $\hat{k} \in \Omega(\kappa)$, there are $a, b \in [\hat{k}]$ such that*

- *For any $r \notin [a+s, b]$, then $E_{\hat{k},r}^{n'_R}$ and $E_{\hat{k},r-s}^{n'_R}$ are both negligible in κ.*
- *For any $r \in [a, b]$, then it holds $e^{-\epsilon/3} \le \frac{E_{\hat{k},r}^{n'_R}}{E_{\hat{k},r-s}^{n'_R}} \le e^{\epsilon/3}$.*

8.3 DDP over Distribution with Geometric Collision

Additional notation. In our analysis, we focus on only 1/3-fraction of high min-entropy elements of the secret set R to deal with the leakage scenario, and consider the Geometric Collision Property only for this subset of elements, which we denote by R'.

Consider any min-hash graph $G = (\mathcal{X}, \mathcal{Y}, \mathcal{E})$. For any set T of iterations of size $\hat{k}$ and any integer r, let $I_{R',T,r}$ be the indicator random variable that is set to 1 if set R' achieves total noise r among the iterations in T.

In this section, we consider a distribution $\tilde{\mathcal{D}}$ of the secret set R'. We define the following measure

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

Geometric collision property. Observe that $D_{T,r}$ corresponds to *the probability (over $\mathcal{D}$) that R' contributes to the noise pattern r*. We would like to show the following:

> For any fixed 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_2 in the above can be slightly relaxed so that the condition holds with all but small probability over the choice of H_2. We observe that the above condition can be captured by showing that $D_{T,r}$ is close to its mean $E^{n'_R}_{\hat{k},r}$ (and then taking advantage of the property of $E^{n'_R}_{\hat{k},r}$ described in Corollary 1). This is essentially to show that $D_{T,r}$ is concentrated around its mean. We could try to apply Chernoff bound to show the concentration property, but we cannot 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. Towards this goal, we introduce a notion called geometric collision property.

Definition 7 (Geometric Collision Property). *Let $\widetilde{\mathcal{D}}$ be a distribution over sets of size n'_R. We say that $\widetilde{\mathcal{D}}$ has the* Geometric Collision Property *if for all $z \in [n'_R]$*

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

Based on this property, we show the following lemma.

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

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

8.4 Distribution with the Geometric Collision Property

Recall that the receiver's input is chosen from the following distribution:

- For each $i \in (n_I, n_B]$, choose x^B_i uniformly at random from the universe U_i. We assume that for $i \neq j$, U_i and U_j are disjoint with the same cardinality $|U_i| = |U_j| := \ell$.

Indeed, the above distribution has the geometric collision property as long as $\ell \geq n_R \cdot \sqrt{n_R}$ (recall $n_R = n_B - n_I$). When considering two random sets R'_0 and R'_1 of size $n'_R = n_R/3$ where their elements are from the above distribution, each position i will have collision with probability $1/\ell$, so we have

$$\Pr[|R'_0 \cap R'_1| = z] = \binom{n'_R}{z} \left(\frac{1}{\ell}\right)^z \cdot \left(\frac{\ell-1}{\ell}\right)^{n'_R - z} \leq \left(\frac{e \cdot n'_R}{z\ell}\right)^z \leq \left(\frac{1}{\sqrt{n_R}}\right)^z .$$

The main issue is that we need to deal with *the leakage stemming from the fact that the hash functions are public*. Therefore, we need to consider a more general class of distributions that captures the leaked version of the above distribution and show that they still possess the geometric collision property.

For brevity, we omit the subscript R and denote $n = n_R$ and $n' = n'_R$. Applying the strong chain rule, we can show that the aforementioned distribution still contains n' blocks each of which maintains high min-entropy, even after the leakage is considered.

Lemma 8. *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 of good 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. Let $\mathcal{D}_{\mathsf{leak}}$ be a distribution over R that has min-entropy at least $\frac{8n}{9}\lg(\ell) + n$. Then, there is a spoiling leakage function f'_G such that with all but negligible probability over $R \sim \mathcal{D}_{\mathsf{leak}}$, the distribution $\widetilde{\mathcal{D}} := \mathcal{D}_{\mathsf{leak}} \mid f'_G(R)$ has the following property:*

> *There exists a set $\{i_1, \ldots, i_{n'}\}$ of size n' such that for $v \in [n']$, and any $(r_{i_1}, \ldots, r_{i_v - 1})$ in the support of $\widetilde{\mathcal{D}}([i_v - 1])$, the random variable $R_{i_v} \sim \widetilde{\mathcal{D}}([i_v]) \mid R_{i_1} = r_{i_1}, \ldots, R_{i_v - 1} = r_{i_v - 1}$ has min-entropy at least $1.5\lg(n)$.*

Given the properties of $\widetilde{\mathcal{D}}$ stated in Lemma 8, by taking advantage of the n' high min-entropy blocks, we show that $\widetilde{\mathcal{D}}$ has the Geometric Collision Property.

Lemma 9. *Assume $\widetilde{\mathcal{D}}$ has the following properties (from the conclusion of Lemma 8):*

- *There exists a set $\{i_1, \ldots, i_{n'}\}$ of size n' such that for $v \in [n']$, and any $(r_{i_1}, \ldots, r_{i_{v-1}})$ in the support of $\widetilde{\mathcal{D}}$ (again we abuse notation and consider $\widetilde{\mathcal{D}}$ to be a distribution over streams $R' = R_{i_1}, \ldots, R_{i_{n'}}$), the random variable $R_{i_v} \sim \widetilde{\mathcal{D}} \mid R_{i_1} = r_{i_1}, \ldots, R_{i_{v-1}} = r_{i_{v-1}}$ has min-entropy at least $1.5\lg(n)$.*

Then $\widetilde{\mathcal{D}}$ (viewed as a distribution over sets) has the Geometric Collision Property (see Definition 7).

Full proofs of Lemma 8 and 9 are deferred to Appendix F.

9 Empirical Evaluation

We conduct empirical evaluations in the public min-hash setting with high individual min-entropy to determine the proper parameter ranges for privacy.

Let k denote the number iterations used in the min-hash (MH) protocol, JI be the Jaccard Index, and (ϵ, δ) be the privacy parameters. Throughout the experiments, we use $n_A = n_B = 10^6$.

Relations between (ϵ, δ), JI and k. In Fig. 6, the left graph shows how the privacy parameters change with respect to JI when $k = 2000$. The right graph in Fig. 6 illustrates the variations in k relative to JI across different privacy parameter settings. These graphs highlight the following observations:

- Keeping k fixed, the privacy parameters become optimal when JI is around 0.5.
- Keeping (ϵ, δ) fixed, the protocol requires the minimum k when JI is around 0.5.
- Keeping JI fixed, higher values of k correspond to improved privacy parameters.

The first two observations are due to the likelihood that a hash function is θ-good being maximized when striking a balance between the two conditions stipulated in Definition 6: (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. For the last observation, note that as the number k of total iterations 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 forming a smoother Poisson Binomial distribution with better privacy guarantee.

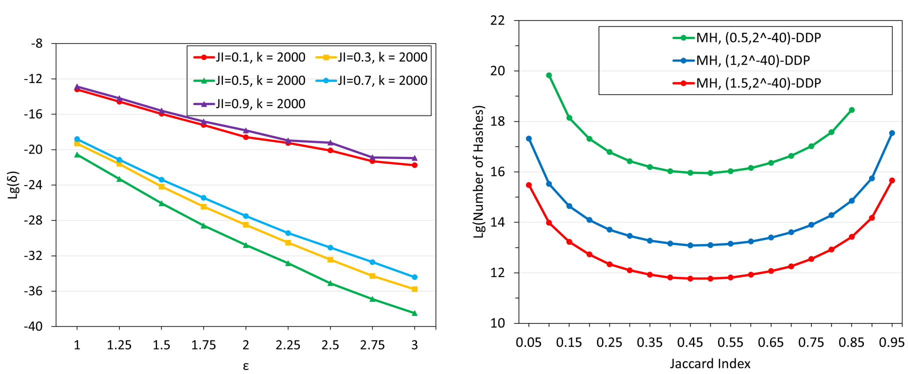

Fig. 6: Relations between (ϵ, δ), JI, and k

We also conduct a similar evaluation in the private min-hash setting, and the results are deferred to Appendix G.

Comparison with Sketch-Flip-Merge (SFM) [37]. We conduct a comparison between min-hash protocols and the current state-of-the-art approach [37], for differentially-private cardinality estimation of set union and intersection. In the absence of available code, we rely on their analysis of the relative root mean squared error (RRMSE) of cardinality estimation, which we describe below.

While our main focus is on comparing the accuracy of Jaccard Index estimation, we encountered challenges in evaluating the accuracy of the Jaccard Index for SFM. 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 [37], 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

$$\mathrm{RRMSE}(\hat{n}_{U,1}, \ldots, \hat{n}_{U,m}; n_U) = \frac{1}{n_U}\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{n}_{U,i} - n_U)^2}.$$

We note that the two protocols have different characteristics.

- The accuracy of SFM depends on its sketch size and the privacy parameter ϵ.
- The accuracy of the min-hash depends on the number k of iterations, which in turn depends on JI and (ϵ, δ).

To accommodate these differences, we first fix the JI and ϵ to compute k for the min-hash, and then we use the same ϵ for SFM while the sketch size of SFM is matched with k. The communication complexity for the two-party computation in the min-hash protocol was calculated according to the PSI-CA protocol [17,62], which exchanges $3k$ objects where each object can be represented with at most 256 bits (either a hash value or an elliptic curve point), which results in $768k$ bits. Therefore, the sketch of SFM is set to be $768k$ bits long; it is a $B \times P$-bit matrix with $P = 24$ and $B = \frac{768k}{P}$. We use $\delta = 2^{-40}$ to mitigate the fact that SFM realizes pure ϵ-DP.

In Figure 7, we demonstrate the comparison of the min-hash (MH) and the SFM. We pair our performance with that of SFM by the privacy guarantee achieved and the Jaccard index. The result shows that our relative error is around 5x smaller than SFM when $\epsilon = 0.5$. For our protocol, better privacy parameters require more hash functions, which also reduce the variance of the JI output, giving better accuracy. On the other hand, SFM relies on adding noise to achieve DP, which means the better the privacy guarantee the less accurate the output will be.

Nevertheless, we admit that SFM gives a more uniform and standard privacy guarantee that does not rely on the Jaccard Index, and provides better flexibility in fine-tuning the tradeoff between privacy and accuracy.

In concluding remarks, it is noteworthy that the SFM 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.[5] When deciding which scheme to use, depending on the specific use case, this observation may need to be taken into account.

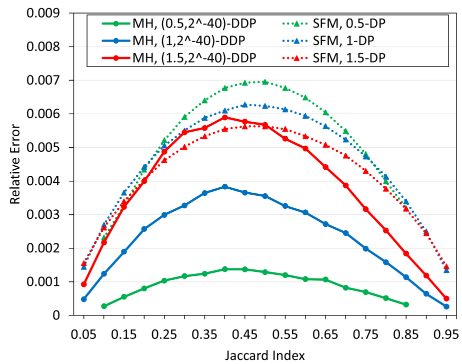

Fig. 7: Accuracy comparison between MH and SFM.

[5] The minimal amount of leakage in the protocol is actually crucial to maintain high min-entropy in the honest input during future min-hash protocol executions.

Acknowledgements

Seung Geol Choi is supported in part by NSF grant CNS-1955319; Dana Dachman-Soled is supported in part by NSF grants CNS-2154705, CNS-1933033, and IIS-2147276; Arkady Yerukhimovich is supported in part by NSF grants CNS-1955620, and CNS-2144798(CAREER).

References

1. Martin Aumüller, Anders Bourgeat, and Jana Schmurr. Differentially private sketches for jaccard similarity estimation. In Shin'ichi Satoh, Lucia Vadicamo, Arthur Zimek, Fabio Carrara, Ilaria Bartolini, Martin Aumüller, Björn Þór Jónsson, and Rasmus Pagh, editors, *Similarity Search and Applications - 13th International Conference, SISAP 2020, Copenhagen, Denmark, September 30 - October 2, 2020, Proceedings*, volume 12440 of *Lecture Notes in Computer Science*, pages 18–32. Springer, 2020.
2. Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *CoRR*, abs/1807.01647, 2018.
3. 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, Heidelberg, Germany.
4. 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.
5. Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Guha Thakurta. Practical locally private heavy hitters. *Advances in Neural Information Processing Systems*, 30, 2017.
6. 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.
7. Amos Beimel, Kobbi Nissim, and Eran Omri. Distributed private data analysis: Simultaneously solving how and what. In David Wagner, editor, *Advances in Cryptology – CRYPTO 2008*, volume 5157 of *Lecture Notes in Computer Science*, pages 451–468, Santa Barbara, CA, USA, August 17–21, 2008. Springer, Heidelberg, Germany.
8. Amos Beimel, Kobbi Nissim, and Eran Omri. Distributed private data analysis: Simultaneously solving how and what. In *Crypto 2008*, pages 451–468. Springer, 2008.
9. 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.
10. 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.

11. A. Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences, International Conference on*, page 21. IEEE Computer Society, 1997.
12. 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.
13. 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.
14. Wei-Ning Chen, Ayfer Ozgur, and Peter Kairouz. The poisson binomial mechanism for secure and private federated learning. ICML, 2022.
15. 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.
16. Emiliano De Cristofaro, Sky Faber, Paolo Gasti, and Gene Tsudik. Genodroid: are privacy-preserving genomic tests ready for prime time? In Ting Yu and Nikita Borisov, editors, *Proceedings of the 11th annual ACM Workshop on Privacy in the Electronic Society, WPES 2012, Raleigh, NC, USA, October 15, 2012*, pages 97–108. ACM, 2012.
17. Emiliano De Cristofaro, Paolo Gasti, and Gene Tsudik. Fast and private computation of cardinality of set intersection and union. In Josef Pieprzyk, Ahmad-Reza Sadeghi, and Mark Manulis, editors, *Cryptology and Network Security, 11th International Conference, CANS 2012, Darmstadt, Germany, December 12-14, 2012. Proceedings*, volume 7712, pages 218–231. Springer, 2012.
18. Dorothy E Denning. Secure statistical databases with random sample queries. *ACM Transactions on Database Systems (TODS)*, 5(3):291–315, 1980.
19. Charlie Dickens, Justin Thaler, and Daniel Ting. Order-invariant cardinality estimators are differentially private. *Advances in Neural Information Processing Systems*, 35:15204–15216, 2022.
20. 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.
21. Benjamin Doerr. Probabilistic tools for the analysis of randomized optimization heuristics. *CoRR*, abs/1801.06733, 2018.
22. 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.
23. Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Serge Vaudenay, editor, *Advances in Cryptology – EUROCRYPT 2006*, volume 4004 of *Lecture Notes in Computer Science*, pages 486–503, St. Petersburg, Russia, May 28 – June 1, 2006. Springer, Heidelberg, Germany.
24. Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.
25. 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.
26. Stefan Dziembowski, Tomasz Kazana, and Maciej Zdanowicz. Quasi chain rule for min-entropy. *Inf. Process. Lett.*, 134:62–66, 2018.

27. Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.
28. Sky Faber. Variants of privacy preserving set intersection and their practical applications. *PhD Thesis*, 2016.
29. 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, Heidelberg, Germany.
30. Paul Francis, Sebastian Probst Eide, and Reinhard Munz. Diffix: High-utility database anonymization. In *Privacy Technologies and Policy: 5th Annual Privacy Forum, APF 2017, Vienna, Austria, June 7-8, 2017, Revised Selected Papers 5*, pages 141–158. Springer, 2017.
31. Andrea Gadotti, Florimond Houssiau, Luc Rocher, Benjamin Livshits, and Yves-Alexandre De Montjoye. When the signal is in the noise: Exploiting diffix's sticky noise. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1081–1098, 2019.
32. Craig Gentry. Fully homomorphic encryption using ideal lattices. In Michael Mitzenmacher, editor, *41st Annual ACM Symposium on Theory of Computing*, pages 169–178, Bethesda, MD, USA, May 31 – June 2, 2009. ACM Press.
33. 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, Heidelberg, Germany.
34. 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.
35. 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.
36. 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.
37. Jonathan Hehir, Daniel Ting, and Graham Cormode. Sketch-flip-merge: Mergeable sketches for private distinct counting. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 12846–12865. PMLR, 2023.
38. Ziyue Huang, Yuan Qiu, Ke Yi, and Graham Cormode. Frequency estimation under multiparty differential privacy: One-shot and streaming. *arXiv preprint arXiv:2104.01808*, 2021.

39. P. Jaccard. Etude comparative de la distribution florale dans une portion des alpes et du jura. 1901.
40. Bo Jiang, Hamid Krim, Tianfu Wu, and Derya Cansever. Refining self-supervised learning in imaging: Beyond linear metric. In *2022 IEEE International Conference on Image Processing, ICIP 2022, Bordeaux, France, 16-19 October 2022*, pages 76–80. IEEE, 2022.
41. Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.
42. Benjamin Kreuter, Craig William Wright, Evgeny Sergeevich Skvortsov, Raimundo Mirisola, and Yao Wang. Privacy-preserving secure cardinality and frequency estimation. 2020.
43. Ping Li, Art B. Owen, and Cun-Hui Zhang. One permutation hashing. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 3122–3130, 2012.
44. 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.
45. Yehuda Lindell and Benny Pinkas. A proof of security of Yao's protocol for two-party computation. *Journal of Cryptology*, 22(2):161–188, April 2009.
46. 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.
47. 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.
48. Luca Melis, George Danezis, and Emiliano De Cristofaro. Efficient private statistics with succinct sketches. *arXiv preprint arXiv:1508.06110*, 2015.
49. Darakhshan Mir, Shan Muthukrishnan, Aleksandar Nikolov, and Rebecca N Wright. Pan-private algorithms via statistics on sketches. In *Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 37–48, 2011.
50. Ilya Mironov, Omkant Pandey, Omer Reingold, and Salil P. Vadhan. Computational differential privacy. In Shai Halevi, editor, *Advances in Cryptology – CRYPTO 2009*, volume 5677 of *Lecture Notes in Computer Science*, pages 126–142, Santa Barbara, CA, USA, August 16–20, 2009. Springer, Heidelberg, Germany.
51. 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, Trondheim, Norway, September 18–20, 2019, Revised Selected Papers 18*, pages 86–98. Springer, 2020.
52. Rasmus Pagh and Nina Mesing Stausholm. Efficient differentially private f_0 linear sketching. *arXiv preprint arXiv:2001.11932*, 2020.
53. Rasmus Pagh and Mikkel Thorup. Improved utility analysis of private countsketch. *Advances in Neural Information Processing Systems*, 35:25631–25643, 2022.

54. 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. `https://eprint.iacr.org/2019/1222`.
55. Igal Sason and Sergio Verdú. Bounds among f-divergences. *CoRR*, abs/1508.00335, 2015.
56. 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.
57. 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.
58. Maciej Skorski. Strong chain rules for min-entropy under few bits spoiled. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1122–1126, 2019.
59. Adam Smith, Shuang Song, and Abhradeep Guha Thakurta. The flajolet-martin sketch itself preserves differential privacy: Private counting with minimal space. *Advances in Neural Information Processing Systems*, 33:19561–19572, 2020.
60. 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.
61. 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.
62. Yang Tan and Bo Lv. Breaking two psi-ca protocols in polynomial time. Cryptology ePrint Archive, Paper 2023/1706, 2023. `https://eprint.iacr.org/2023/1706`.
63. Chayant Tantipathananandh, Tanya Y. Berger-Wolf, and David Kempe. A framework for community identification in dynamic social networks. In Pavel Berkhin, Rich Caruana, and Xindong Wu, editors, *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, pages 717–726. ACM, 2007.
64. Lun Wang, Iosif Pinelis, and Dawn Song. Differentially private fractional frequency moments estimation with polylogarithmic space. In *International Conference on Learning Representations*, 2022.
65. 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 Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4816–4823. ijcai.org, 2019.
66. Ziqi Yan, Jiqiang Liu, Gang Li, Zhen Han, and Shuo Qiu. Privmin: Differentially private minhash for jaccard similarity computation. *CoRR*, abs/1705.07258, 2017.
67. 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.
68. Andrew Chi-Chih Yao. How to generate and exchange secrets (extended abstract). In *27th Annual Symposium on Foundations of Computer Science*, pages 162–167, Toronto, Ontario, Canada, October 27–29, 1986. IEEE Computer Society Press.
69. Fuheng Zhao, Dan Qiao, Rachel Redberg, Divyakant Agrawal, Amr El Abbadi, and Yu-Xiang Wang. Differentially private linear sketches: Efficient implementations and applications. *Advances in Neural Information Processing Systems*, 35:12691–12704, 2022.
70. Ferdinand Österreicher. Csiszar's f-divergence-basic properties. *RGMIA Research Report Collection*, 2002.

A Hockey stick divergence

We first review hockey stick divergence [55], also known as elementary divergences [70] or α-distance [3]. In this paper, we only consider discrete sample space, although many of arguments naturally extend to the continuous space.

Definition 8. *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-stack divergence.

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

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

Therefore, the hockey-stick divergence captures the inequalities between output distributions from neighboring inputs. We use hockey-stick divergence to analyze the privacy loss.

The hockey-stick divergence as a f-divergence. It is known that the hockey-stick divergence is a kind of f-divergence [3]. Therefore, the hockey-stick divergence satisfies the joint convexity and data processing inequality [3,2].

Lemma 10 (Joint convexity). *For all $0 \leq \lambda \leq 1$ and $\alpha \geq 1$, any probability measures X_1, X_2, Y_1, Y_2 satisfy*

$$\mathsf{D}^{\mathsf{hs}}_{\alpha}(\lambda X_1 + (1-\lambda)X_2, \lambda Y_1 + (1-\lambda)Y_2) \leq \lambda \mathsf{D}^{\mathsf{hs}}_{\alpha}(X_1, Y_1) + (1-\lambda)\mathsf{D}^{\mathsf{hs}}_{\alpha}(X_2, Y_2).$$

The data processing inequality property guarantees that post-processing cannot increase the divergence.

Lemma 11 (Data processing inequality.). *For any probability measures X, Y over Z and any function g with domain Z, we have*

$$\mathsf{D}^{\mathsf{hs}}_{\alpha}(g(X), g(Y)) \leq \mathsf{D}^{\mathsf{hs}}_{\alpha}(X, Y).$$

B Proof of Lemma 3

We set $a = \frac{np+s(1-p)\cdot e^{\epsilon/s}}{e^{\epsilon/s}\cdot(1-p)+p}$; we chose a so that $\frac{n-a}{a-s} = e^{\epsilon/s} \cdot \frac{1-p}{p}$ in order to work out the following:

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

We also set $b = \frac{e^{\epsilon/s}\cdot np}{(1-p)+e^{\epsilon/s}\cdot p}$; we chose b so that $\frac{n-b}{b} = e^{-\epsilon/s} \cdot \frac{1-p}{p}$ in order to work out the following:

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

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

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

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

To see the asymptotic measure of d, let $t = s(1-p) \cdot e^{\epsilon/s}$ and then we have $a = \frac{\mu+t}{e^{\epsilon/s}\cdot(1-p)+p}$; Then, we have

$$
\begin{aligned}
d &= \frac{\mu - a}{\mu} - \frac{s}{\mu} \\
&= \frac{\mu(e^{\epsilon/s}\cdot(1-p)+p) - \mu - t}{\mu \cdot (e^{\epsilon/s}\cdot(1-p)+p)} - \frac{s}{\mu} \\
&\geq \frac{\mu((1+\epsilon/s)\cdot(1-p)+p) - \mu - t}{\mu \cdot e^{\epsilon/s}} - \frac{s}{\mu} \\
&= \frac{(\epsilon/s)\cdot(1-p)}{e^{\epsilon/s}} - \frac{t}{\mu \cdot e^{\epsilon/s}} - \frac{s}{\mu} \\
&\geq (\epsilon/s) \cdot \frac{(1-p)}{e^{\epsilon}} - \frac{t+s}{\mu} \\
&= \Theta(1/\lg\lg\kappa) - \tilde{O}(1/\kappa)
\end{aligned}
$$

Since we have $d = \Omega(\frac{1}{\lg\lg\kappa})$ and $\mu \in \Theta(\kappa)$, $\Pr_{X\leftarrow \mathrm{B}(n,p)}[X \leq a+s]$ is negligible in κ.

C 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

$$
\begin{aligned}
&\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}(P_C, P_{C+1}) \\
&\quad \leq \max\left(\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{\mathrm{B}(\lceil \frac{n}{2}\rceil, \eta_{+\theta})}, P_{\mathrm{B}(\lceil \frac{n}{2}\rceil, \eta_{+\theta})+1}\right), \mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{\mathrm{B}(\lceil \frac{n}{2}\rceil, \eta_{-\theta})}, P_{\mathrm{B}(\lceil \frac{n}{2}\rceil, \eta_{-\theta})+1}\right)\right).
\end{aligned}
$$

To show the above, we follow a similar structure to the proof of [14, Theorem 3.3], which analyzes the Renyi divergence of a slightly different version of (non-additive) Poisson binomial mechanism. We also rely on the joint convexity[6] (Lemma 10) and data processing inequality (Lemma 11) to upper bound the hockey-stick divergence of two Poisson binomial distributions with that of two binomial distributions.

We start with showing that the upper bound of the hockey-stick divergence is reached at extreme points. The proof is similar to [14, Lemma 3.5].

Lemma 12.

$$
\begin{aligned}
&\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}(P_C, P_{C+1}) && (2) \\
&\quad \leq \max_{j\in[n]} \mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{\mathrm{B}(j,\eta_{-\theta})+\mathrm{B}(n-j,\eta_{+\theta})}, P_{\mathrm{B}(j,\eta_{-\theta})+\mathrm{B}(n-j,\eta_{+\theta})+1}\right), && (3)
\end{aligned}
$$

[6] [14] actually uses joint quasi-convexity, which is implied by joint convexity but not vice versa.

Proof. Note that there is $\lambda \in [0,1]$ such that $p_n = \lambda\left(\eta_{-\theta}\right) + (1-\lambda)\left(\eta_{+\theta}\right)$. Let $C_{-n} = \sum_{j=1}^{n-1} c_j$. Denote the convolution $\mathsf{Conv}(a, a', b) = a \cdot (1-b) + a' \cdot b$. Then, we have

$$\begin{aligned}
\Pr[C=\ell] &= \mathsf{Conv}(\Pr[C_{-n}=\ell], \Pr[C_{-n}=\ell-1], p_n) \\
&= \mathsf{Conv}\left(\Pr[C_{-n}=\ell], \Pr[C_{-n}=\ell-1], \lambda\cdot\eta_{-\theta} + (1-\lambda)\cdot\eta_{+\theta}\right) \\
&= \lambda\mathsf{Conv}(\Pr[C_{-n}=\ell], \Pr[C_{-n}=\ell-1], \eta_{-\theta}) \\
&\quad + (1-\lambda)\mathsf{Conv}(\Pr[C_{-n}=\ell], \Pr[C_{-n}=\ell-1], \eta_{+\theta}) \\
&= \lambda\Pr[C_{-n} + \textsc{Ber}(\eta_{-\theta}) = \ell] + (1-\lambda)\Pr[C_{-n} + \textsc{Ber}(\eta_{+\theta}) = \ell]
\end{aligned}$$

In other words, we have

$$P_C = \lambda P_{C_{-n}+\textsc{Ber}(\eta_{-\theta})} + (1-\lambda) P_{C_{-n}+\textsc{Ber}(\eta_{+\theta})}.$$

Now using Lemma 10, we have

$$\begin{aligned}
&\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_C, P_{C+1}\right) \\
&\leq \lambda\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{C_{-n}+\textsc{Ber}(\eta_{-\theta})}, P_{C_{-n}+\textsc{Ber}(\eta_{-\theta})+1}\right) + (1-\lambda)\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{C_{-n}+\textsc{Ber}(\eta_{+\theta})}, P_{C_{-n}+\textsc{Ber}(\eta_{+\theta})+1}\right) \\
&\leq \max\left\{\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{C_{-n}+\textsc{Ber}(\eta_{-\theta})}, P_{C_{-n}+\textsc{Ber}(\eta_{-\theta})+1}\right), \mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{C_{-n}+\textsc{Ber}(\eta_{+\theta})}, P_{C_{-n}+\textsc{Ber}(\eta_{+\theta})+1}\right)\right\}
\end{aligned}$$

Repetitively applying the joint convexity on the remaining $n-1$ random variables yields the following

$$\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_Y, P_{Y+1}\right) \leq \max_{j\in[n]} \mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{N_j}, P_{N_j+1}\right),$$

where $N_j \sim \mathrm{B}(j, \eta_{-\theta}) + \mathrm{B}(n-j, \eta_{+\theta})$.

Next, we apply data processing inequality to simplify (3) from above lemma.

Lemma 13. *(3) is upper bounded by*

$$\max\left(\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{+\theta})}, P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{+\theta})+1}\right), \mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{-\theta})}, P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{-\theta})+1}\right)\right) \quad (4)$$

Proof. The proof is similar to [14, Lemma 3.6]. We apply the data processing inequality. More specifically, consider two distributions X, Y and then adding a binomial random variable Z as the post-processing step. Then, the addition of Z doesn't increase the divergence. In other words,

$$\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_X, P_Y\right) \geq \mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{X+Z}, P_{Y+Z}\right).$$

Now, consider (3) and let j^* be the index that leads to the maximum.

- If $j^* \leq n/2$, we have

$$\begin{aligned}
&\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{+\theta})}, P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{+\theta})+1}\right) \\
&\quad \geq \mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{\mathrm{B}(j^*,\eta_{-\theta})+\mathrm{B}(n-j^*,\eta_{+\theta})}, P_{\mathrm{B}(j^*,\eta_{-\theta})+\mathrm{B}(n-j^*,\eta_{+\theta})+1}\right).
\end{aligned}$$

  This is because $n/2 \leq n - j^*$. In this case, $Z = \mathrm{B}(j^*, \eta_{-\theta}) + \mathrm{B}(n - j^* - \lceil\frac{n}{2}\rceil, \eta_{+\theta})$.

– Likewise, if $j^* \geq n/2$, we have

$$\begin{aligned}&\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{-\theta})}, P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{-\theta})+1}\right)\\ &\quad\geq \mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{\mathrm{B}(j^*,\eta_{-\theta})+\mathrm{B}(n-j^*,\eta_{+\theta})}, P_{\mathrm{B}(j^*,\eta_{-\theta})+\mathrm{B}(n-j^*,\eta_{+\theta})+1}\right).\end{aligned}$$

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 $\epsilon > 0$, we have*

$$\begin{aligned}&\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_C, P_{C+s}\right)\\ &\leq \max\left(\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{+\theta})}, P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{+\theta})+s}\right), \mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{-\theta})}, P_{\mathrm{B}(\lceil\frac{n}{2}\rceil,\eta_{-\theta})+s}\right)\right)\end{aligned}$$

Finally, to give a bound on the divergence, we can apply Lemma 3 to argue 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 $\mathsf{D}^{\mathsf{hs}}_{e^\epsilon}\left(P_{C+s}, P_C\right) \leq \mathsf{negl}(\kappa)$. □

D Proof of Theorem 3

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 $1/n_R$, where n_A, n_B, n_I, n_R, n'_R are defined above. In particular,

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

Further, we require $\theta \leq 1/10$.

Bundle of good iterations K_θ. The total number of iterations in the min-hash protocol π_{PH} 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 $k_g = |K_\theta|$. Of these k_g number of θ-good iterations (except in S_{x^*}), we further divide them 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 $K_{bad} := \overline{K_\theta} = \overline{G_\theta} \cup S_{x^*}$.

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}:

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

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

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

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

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

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 **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. We first consider the original distribution on P_2's secret set R to be 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 $f(R) = \gamma$ of length L. L is set 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 $Z \subseteq R$, 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. 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

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

Intuitively, no elements in the fixed set Z can be minimum hash value in the jth iteration. In this sense, the iteration j is still available for other elements than those in Z to become the winner of delivering the minimum hash value.

Existence of a good bundle. We now describe a process for identifying a "good" bundle of iterations in the sense that given the fixed hash, the distribution $\mathcal{D}$ (after the leakage) satisfies the DP-like property. We show that with all but negligible probability, process **IsAGoodBundle** succeeds on at least one bundle $K_{\theta,i}$ where $i \in [u]$.

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

1. Consider $G \leftarrow \mathbf{MinhashG}_{H_1}(A, I, x^*, H_2)$.
2. Let $\mathcal{D}_{1,i} := \mathcal{D} \mid \mathbf{out}_{-i}$. In other words, $\mathcal{D}_{1,i}$ is the distribution $\mathcal{D}$ on R, but conditioned on the output vector $\mathbf{out}_{-i}$. 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.
   We note that by applying [20, Lemma 2.2], the average min-entropy of $\mathcal{D}|\mathbf{out}_{-i}$ is at least $n_R \lg \ell - L - \kappa$, which implies that the min-entropy of $\mathcal{D}|\mathbf{out}_{-i}$ is at least $n_R \lg \ell - L - 2\kappa \geq \frac{8n_R}{9} \lg \ell + n_R$ with probability $1 - 2^{-\kappa}$ (assuming that $n_R \geq 2\kappa$). Therefore, the probability that the process outputs $\mathsf{FAIL}_{1,i}$ is at most $2^{-\kappa} \leq \mathsf{negl}(\kappa)$.
3. Due to Lemmas 8 and 9, there is a leakage function $f_G(R)$ which leaks $V = \{j_1, \ldots, j_{n'_R}\}$ such that with all but negligible probability 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. The leakage function f also leaks $T = \mathsf{Avail}_G(K_{\theta,i}, R \setminus R')$. Let $\mathcal{D}_{2,i} := \mathcal{D}_{1,i} | f_G(R)$.
4. If it holds $|T| \leq \frac{1}{10}|K_{\theta,i}|$, output $\mathsf{FAIL}_{3,i}$ and terminate. Let $k_v = |T|$.
5. Using G constructed above, compute
$$D_{T,r}(\mathcal{D}_{2,i}) = \Pr_{R' \sim \mathcal{D}_{2,i}}[I_{R',T,r}].$$
   Check if there are a and b satisfying the following conditions:
   - For $r \notin [a + \lg\lg\kappa, b]$, $D_{T,r}(\mathcal{D}_{2,i})$ is negligible in κ.
   - For $r \in [a, b]$, it holds that $e^{-\epsilon/3} E_{k_v,r}^{n'_R} \leq D_{T,r}(\tilde{\mathcal{D}}) \leq e^{\epsilon/3} E_{k_v,r}^{n'_R}$.

   Output $\mathsf{FAIL}_{4,i}$ and terminate, if the above check fails.
6. Output SUCCESS.

SUCCESS with high probability. We already argued that $\mathsf{FAIL}_{1,i}$ and $\mathsf{FAIL}_{2,i}$ occur with negligible probability. We now argue that $\mathsf{FAIL}_{3,i}$ occurs with negligible probability. Recall $k_b = |K_{\theta,i}|$.

Lemma 14. *Fix H_1, A, I, x^*, and consider choosing H_2 and constructing $G \leftarrow$ $\mathbf{MinhashG}_{H_1}(A, I, x^*, H_2)$. Then, it holds that*
$$\Pr_{H_2}\left[|T| \leq \frac{k_b}{10}\right] \leq \mathsf{negl}(\kappa).$$

Proof. 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 $p_j \leq 1-(\eta_{-\theta})^{1/n'}$ and $\Pr[X_j = 1] = 1-(1-p_j)^{n-n'} \leq 1-(\eta_{-\theta})^{\frac{n-n'}{n'}} = 1 - (\eta_{-\theta})^2 \leq 1 - (2/5)^2$. Therefore, we have
$$m := \mathbf{E}\left[\sum_{j=1}^{k_b} X_j\right] \leq k_b \cdot (1 - (\eta_{-\theta})^2) \leq 0.84 k_b$$

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

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

We now analyze $\mathsf{FAIL}_{4,i}$. In particular, 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

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

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

Further, 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 $(\mathbf{out}_{-i}, R')$, 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 process* **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 9 (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 15. *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 $\mathbf{out}_{-i^*}$, 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 Lemma 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

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

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_{PH}. 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 $\mathbf{out}_{-i^*}$ 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 $out^+_{bad} - out_{bad} \leq s = \lg\lg(\kappa)$. 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 15, 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.

E Proof of Lemma 7

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

We first show the following lemma holds.

Lemma 16. *Let $\mathcal{D}$ be a distribution over sets of size n'_R with geometric collision property. Fix H_1 and consider $\hat{k}, \theta, a, b$ specified in Lemma 1 with the same requirements. Then, we have the following:*

Case 1: *If $r \notin [a+s, b]$, then we have*

$$\Pr_{H_2}\left[D_{\hat{k},r}(\tilde{\mathcal{D}}) \leq \mathsf{negl}(\kappa)\right] \geq 1 - \mathsf{negl}(\kappa).$$

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

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

Then, Lemma 7 follows by taking a union bound over different cases of r. $\square$

E.1 Proof of Lemma 16

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

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

Recall that $D_r(\tilde{\mathcal{D}}) = \sum_{R'} \rho(R') \cdot I_{R',r}$. 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

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

Note that $\sum_{R' \in \mathsf{Heavy}} \rho(R') \ge 1/p(\kappa)$. Now, since $\tilde{\mathcal{D}}$ and H_2 are independent, the above implies that

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

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 17. *Fix H_1. Consider a graph $G \leftarrow \mathbf{MinhashG}_{H_1}(A, I, x^*, H_2)$. Consider any set T iterations in G such that $|T| = \hat{k}$. 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 (Lemma 17). Let $p = 1 - (\eta_{-\theta})^{1/n'_R}$. We first show that $p \le 1/n'_R$. Recall $\theta \le 1/10$, which implies $e^{-1} \le 1/2 - \theta = \eta_{-\theta}$. Therefore, we have $(1 - 1/n'_R)^{n'_R} \le e^{-1} \le \eta_{-\theta}$, so $1 - 1/n'_R \le (\eta_{-\theta})^{1/n'_R}$. Therefore, we have $1 - (\eta_{-\theta})^{1/n'_R} \le 1/n'_R$.

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

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

Since n'_R is much larger than $\hat{k}$, the above probability becomes negligible in κ.

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

Lemma 18. *Fix H_1. We set the parameters for $\hat{k}, a$ 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*

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

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

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

The second inequality holds due to Lemma 17. It is left to bound $\Pr[I_{X,m}]$ for $m \in (r-\zeta, r]$. We observe that the following holds:

$$\Pr_{H_2}[I_{X,m}] = \Pr[I_{R'_i,m} |\ I_{Z,0}].$$

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

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

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

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

Lemma 19. *We set the parameters for $H_1, \hat{k}, a$ and b as stated in Lemma 7. Let $\widetilde{\mathcal{D}}$ be a distribution with the geometric collision property. Then, for every $a \leq r \leq b$, we have*

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

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

$$
\begin{aligned}
\mathsf{Var}_{H_2}[D_r] \quad &= \sum_{R'_i, R'_j} \rho(R'_i) \cdot \rho(R'_j) \cdot (\mathbb{E}[I_{R'_i,r} \cdot I_{R'_j,r}] - \mathbb{E}[I_{R'_i,r}] \cdot \mathbb{E}[I_{R'_j,r}]) \\
&\le \sum_{R'_i, R'_j : |R'_i \cap R'_j| \ge 1} \rho(R'_i) \cdot \rho(R'_j) \cdot \mathbb{E}[I_{R'_i,r} \cdot I_{R'_j,r}] \\
&= \sum_{z=1}^{n'_R} \Pr_{R'_i, R'_j \sim \widetilde{\mathcal{D}}} [|R'_i \cap R'_j| = z] \cdot \mathbb{E}[I_{R'_i,r} \cdot I_{R'_j,r}] \\
&\le \sum_{z=1}^{n'_R} \left(\frac{1}{\sqrt{n_R}}\right)^z \cdot \left(1 + \frac{\zeta \cdot (e^{\zeta\epsilon/3} + 1)}{\eta_{-\theta}{}^{zk/n'_R}}\right) \cdot \left(E_r^{n'_R}\right)^2 \\
&\le \sum_{z=1}^{n'_R} \left(\frac{1}{\sqrt{n_R}}\right)^z \cdot \left(\zeta \cdot \frac{e^{\zeta} + 2}{(2/5)^{\zeta/3}}\right) \cdot \left(E_r^{n'_R}\right)^2 \\
&\le \sum_{z=1}^{n'_R} \left(\frac{1}{\sqrt{n_R}}\right)^z \cdot \left(8^{\zeta+1}\right) \cdot \left(E_r^{n'_R}\right)^2 \\
&= 8 \cdot \left(E_r^{n'_R}\right)^2 \cdot \sum_{z=1}^{n'_R} \left(\frac{\lg^3 \kappa}{\sqrt{n_R}}\right)^z \\
&\le \frac{16 \lg^3 \kappa}{\sqrt{n_R}} \left(E_r^{n'_R}\right)^2.
\end{aligned}
$$

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 $\tilde{\mathcal{D}}$ and Lemma 18. The third inequality holds with $\epsilon \le 3$ since $\theta < 1/10$ and $\hat{k}$ is much smaller than n'_R. $\square$

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

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

F Distribution with the Geometric Collision Property

In this section, we show that the distribution $\mathcal{D}_{2,i}$ described in process **IsAGoodBundle** possesses the geometric collision property. Note that the receiver's input is chosen from the following distribution:

– For each $i \in (n_I, n_B]$, choose x_i^B uniformly at random from the universe U_i. We assume that for $i \neq j$, U_i and U_j are disjoint with the same cardinality $|U_i| = |U_j| := \ell$.

Indeed, the above distribution has the geometric collision property as long as $\ell \geq n_R \cdot \sqrt{n_R}$ (recall $n_R = n_B - n_I$). When considering two random sets R'_0 and R'_1 of size $n'_R = n_R/3$ where their elements are from the above distribution, each position i will have collision with probability $1/\ell$, so we have

$$\Pr[|R'_0 \cap R'_1| = z] = \binom{n'_R}{z} \left(\frac{1}{\ell}\right)^z \cdot \left(\frac{\ell - 1}{\ell}\right)^{n'_R - z} \leq \left(\frac{e \cdot n'_R}{z\ell}\right)^z \leq \left(\frac{1}{\sqrt{n_R}}\right)^z .$$

The main issue is that we need to deal with *the leakage stemming from the fact that the hash functions are public*. Therefore, we need to consider a more general class of distributions that captures the leaked version of the above distribution and show that they still possess the geometric collision property.

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* [58]; 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 [58] into our setting, in which itself needs a significant amount of modification. 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 5 (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.,*

$$\forall y \in Im(f) \setminus \{\perp\}, \forall r \text{ s.t. } f(r) = y, \forall i \in V :$$
$$\Pr\left[R_i = r_i \;\middle|\; R_{<i} = r_{<i},\ y\right] \leq \frac{2^\delta}{n^{1.5}}.$$

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* $R_W := \{R_i : i \in W\}$.

F.1 Proof of Theorem 5

By following the general idea of [58], 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

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

Therefore, using a union bound across all $i \in [n]$, we have

$$\Pr[r \notin \mathcal{U}'] \leq \epsilon \cdot n.\square$$

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 $\textsc{sp}_i^1(r)$ denote the surprise of the ith element of r. More formally,

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

   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 $\textsc{sp}_i^2(r)$ denote the surprise of the sequences with a similar surprise level in aggregate.

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

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

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

   (d) Otherwise, let $f_i(r) = (\mathsf{r}(\textsc{sp}_i^1(r)), \mathsf{false}, H_i)$ 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

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

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

$$\frac{\Pr[r_i \wedge r_{<i} \wedge y_{>i}]}{\Pr[y_{i,1} \mid r_{<i} \wedge y_{>i}] \Pr[r_{<i} \wedge y_{>i}]} = \frac{\Pr[r_i \mid r_{<i} \wedge y_{>i}]}{\Pr[y_{1,i} \mid r_{<i} \wedge y_{>i}]} = \frac{2^{-\textsc{sp}_i^1(r)}}{2^{-\textsc{sp}_i^2(r)}} \leq \frac{2^\delta}{n^{1.5}}. \quad (5)$$

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 such that

$$\Pr[R_i = r_i | R_{<i} = r_{<i}, y_{i,1}, y_{i,2}, y_{>i}]] \geq 0,$$

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.

F.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 4 stated in Section 7 is restatement of Theorem 5 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 $\rho_i = h(R_i)$.

F.3 Towards Achieving Geometric Collision Property

Remember that 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, in this section, by applying Theorem 5 to this distribution, we show that even with the leakage, there are at least $n_R/3$ elements that preserves enough min-entropy. In the next section, we show how these elements with sufficient min-entropy give the geometric collision property. For brevity, we let $n := n_R$ in this subsection and $n' := n'_R = n_R/3$.

Remark 1 (Getting rid of tiny parts). Similar to [58, 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 5. Additionally, $\forall y \in Im(f')$, we have $\Pr_R[f'(R) = y] \geq \epsilon n/(2|Im(f)|)$.

F.4 Proof of Lemma 8

In Lemma 8, by setting $\ell \geq 4n^3$ and assuming sufficient min-entropy of R, we 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

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

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

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

$$\Pr_{\tilde{\mathcal{D}}}[r] = \Pr_{\mathcal{D}_{\mathsf{leak}}}[r \mid y] = \frac{\Pr_{\mathcal{D}_{\mathsf{leak}}}[r \wedge y]}{\Pr_{\mathcal{D}_{\mathsf{leak}}}[y]} \le \frac{2^{-(\frac{8n}{9} \lg \ell + n)}}{(\epsilon n/2) \cdot 2^{-0.5n \lg n}} \tag{6}$$

$$= 2^{-(\frac{8}{9} \log \ell - 0.5 \lg n + 1) \cdot n} \cdot (2/\epsilon n), \tag{7}$$

which suggests $\tilde{\mathcal{D}}$ has min-entropy at least $(\frac{8}{9} \log \ell - 0.5 \lg n + 1) \cdot n - \lg(2/\epsilon n)$.

We argue that 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 5). We will show that if $|V| \le n/3$ there exists a point r in the support of $\tilde{\mathcal{D}}$ such that $\Pr_{\tilde{\mathcal{D}}}[r] > 2^{-(\frac{8}{9} \log \ell - 0.5 \lg n + 1) \cdot n} \cdot (2/\epsilon n)$, which contradicts the min-entropy of $\tilde{\mathcal{D}}$.

First, find any value r_V^* such that $\Pr_{\tilde{\mathcal{D}}}[R_V = r_V^*] \ge \frac{1}{\ell^{|V|}}$. Note that r_V^* must exist since the support size of R_V is at most $\ell^{|V|}$.

Let $S_W(r_V^*) = \{r : r_V = r_V^* \wedge \Pr_{\tilde{\mathcal{D}}}[R = r] > 0\}$.

$$\Pr_{\tilde{D}}[R \in S_W(r_V^*)] = \Pr[R_V = r_V^*] \ge \frac{1}{\ell^{|V|}}.$$

Second, we show that $|S_W(r_V^*)| \le (2 \cdot n^{1.5})^{|W|}$. Consider any $r \in S_W(r_V^*)$. Applying the fourth condition of Theorem 5 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 $|S_W(r_V^*)|$ 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 $r^* \in S_W(r_V^*)$ for which

$$\Pr_{\tilde{\mathcal{D}}}[R = r^*] \ge \frac{1}{(\ell)^{|V|}} \cdot \frac{1}{(2 \cdot n^{1.5})^{|W|}}.$$

Therefore, we have

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

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

To reach contradiction to (7), 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 (7). We therefore conclude that $|V| \geq n/3$.

F.5 Proof of Lemma 9

Note that we can equivalently view r' in the support of $\widetilde{\mathcal{D}}$ 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 $r', \bar{r}'$ in the support of $\widetilde{\mathcal{D}}$, we have that $|r' \cap \bar{r}'| = z$ 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 $\bar{r}'$ indexed by Z (denoted $\bar{r}'_Z$) 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 $\bar{r}'$ indexed by $[n'] \setminus Z$ (denoted $\bar{r}'_{\overline{Z}}$) are disjoint.

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

$$\begin{aligned}
\Pr_{r',\bar{r}' \leftarrow \widetilde{\mathcal{D}}}[|r' \cap \bar{r}'| = z] &= \sum_{Z \subseteq [n'], |Z| = z} \Pr_{r',\bar{r}' \leftarrow \widetilde{\mathcal{D}}}\left[(r'_Z = \bar{r}'_Z) \wedge \left(r'_{\overline{Z}} \cap \bar{r}'_{\overline{Z}}) = \emptyset\right)\right] \\
&\leq \sum_{Z \subseteq [n'], |Z| = z} \Pr_{r',\bar{r}' \leftarrow \widetilde{\mathcal{D}}}[r'_Z = \bar{r}'_Z] \\
&\leq \sum_{Z \subseteq [n'], |Z| = z} \left(\frac{1}{n^{1.5}}\right)^z \\
&\leq \left(\frac{e \cdot (n/3)}{z}\right)^z \cdot \left(\frac{1}{n^{1.5}}\right)^z \\
&\leq \left(\frac{1}{n^{0.5}}\right)^z .
\end{aligned}$$

G Empirical Evaluation for Private Min-hash

We conduct empirical evaluations in the private min-hash setting to showcase the proper parameter ranges for privacy. In Fig. 8, the left figure shows the trade-off between ϵ and δ with respect to different Jaccard indices n_I and the number of iterations when $n_A = n_B = 10^6$. The right figure shows the number of iterations required to achieve DP for a given Jaccard index for various values of ϵ, δ.

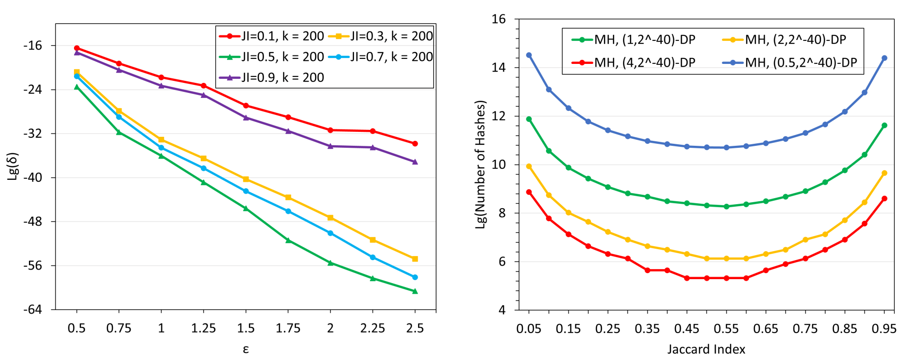

Fig. 8: $n_A = n_B = 10^6$. Private min-hash setting.