Differentially private distance-sensitive set representations

Paul Cesaretti[1]

[1]Department of Computer Science, CUNY Graduate Center

Abstract

We study the problem of representing a set $S \subseteq \{0,1\}^d$ in a differentially private manner to answer distance-sensitive membership queries. This extends two prior lines of work: Patel et al., who give optimal (ε, δ)-DP set representations for exact membership queries via a random linear system framework, and Goswami et al., who give near-optimal space bounds for distance-sensitive filters in Hamming space. We run the Patel et al. mechanism unchanged on the image of S under the Goswami et al. signature σ, and recover distance-sensitivity at query time by expanding each query into an ℓ_1 ball of candidate signatures; since σ is fixed public randomness independent of S, privacy is inherited unchanged. In the regime $r = O(c \log(k/\alpha))$, with c a free approximation factor, we achieve error at most α on both sides, at the cost of a $\Theta(\log(k/\alpha))$-factor increase in the privacy parameter and the dominant space term relative to exact membership. We also identify and formally state several open problems arising from this work, including a unified space lower bound combining the compression arguments of both prior papers, and leave open whether this penalty is necessary, along with a unified space lower bound.

Keywords

differential privacy, distance-sensitive filters, set representations, vector signatures

1. Introduction

Differentially private (DP) data structures are a fundamental primitive in privacy-preserving computation, enabling representations of sensitive datasets that support useful queries while providing formal privacy guarantees. A central problem in this space is the *private set representation* problem: given a set S of size k from a large universe U, construct a compact (ε, δ)-DP encoding of S that supports membership queries with small error probability.

Patel et al. [1] recently gave an essentially complete solution to this problem for *exact membership* queries, achieving error probability $\alpha = 1/(e^{\varepsilon} + 1)$, which matches their information-theoretic lower bound. Their construction embeds S into a random linear system over a finite field, departing from prior noise-injection approaches.

A natural and important extension is to *distance-sensitive* membership queries: given a query $u \in U$, determine whether $D(u, S) \leq r$ (i.e., whether u is within distance r of some element of S). Distance-sensitive filters have been studied in the non-private setting by Goswami et al. [2] and Kirsch–Mitzenmacher [3], with applications to plagiarism detection, network monitoring, and approximate nearest-neighbor search. However, no prior work has studied distance-sensitive filters under differential privacy.

Our contribution We construct the first (ε, δ)-DP distance-sensitive set representation, which we call a *DPDS-Representation*. The two prior frameworks solve complementary halves of the problem but do not compose directly: the Patel et al. linear system provides privacy through a mechanism that is agnostic to what its inputs mean, but its decoder answers only *exact* queries; the Goswami et al. vector signature σ provides the missing distance-sensitivity by mapping Hamming-near pairs to signatures that are close in ℓ_1 and Hamming-far pairs to signatures that are far, but "close" is not "equal," and the linear system has no notion of approximate satisfaction.

We resolve this mismatch on the decode side rather than by modifying the encoding. The encoder is exactly the Patel et al. mechanism applied to the image set $\sigma(S')$ of a subsampled $S' \subseteq S$; in particular the encoding distribution is unchanged, so the Patel et al. privacy proof carries through

ICTCS 2026: Italian Conference on Theoretical Computer Science, September 7–9, 2026, Udine, Italy

pcesaretti@gradcenter.cuny.edu (P. Cesaretti)

verbatim, with $\sigma(s)$ playing the role of s. To answer a query u, the decoder enumerates the ℓ_1 ball $Q_u = \{z : \|z - \sigma(u)\|_1 \leq \Psi'\}$ of candidate signatures around $\sigma(u)$ and accepts if and only if some candidate satisfies its linear constraint. The near guarantee places a near witness's signature in Q_u deterministically (so there are no bin-boundary failures), while the far guarantee keeps every encoded signature out of Q_u with high probability. A key design choice, inherited from the impossibility of one-sided error under DP, is that we accept false negatives as part of the error probability α alongside false positives.

The cost of decode-side expansion is concentrated in the size $N_Q = |Q_u|$ of the candidate ball, which controls the field size $q = \lceil 2N_Q/\alpha \rceil$ and hence the privacy parameter, the space, and the decode time. In the regime $r = O(c \log(k/\alpha))$ one has $N_Q = \mathrm{poly}(k/\alpha)$, and the construction achieves error at most α on both sides with privacy $\varepsilon = \ln(q-1)$. Relative to exact membership this is a $\Theta(\log(k/\alpha))$-factor penalty in ε and in the dominant space term—not a lower-order correction—and it grows as the approximation factor $c \to 1^+$. Crucially, c remains a free user parameter throughout, decoupled from the error rate α; this is what the decode-side approach buys over alternatives that either couple c to α or degrade the privacy–utility tradeoff to polynomial. Outside the radius regime, or as $c \to 1^+$, the cost approaches that of a naive Bloom-filter ball query, which Goswami et al. identify as the baseline.

Organisation Section 2 gives a background of the related works in the three research areas our work extends, as well as a description of what these works lack for our construction. Section 3 recalls the two prior frameworks needed in our construction. Section 4 presents our construction. Section 5 gives the error, privacy, and space analysis. Section 6 states the main open problems arising from this work.

2. Related work

Our work sits at the intersection of three research threads: succinct probabilistic data structures (Bloom filters and their extensions), differentially private linear sketches, and private set representations. We survey each in turn and explain how Patel et al. [1] and Goswami et al. [2] fit into this landscape.

2.1. Bloom filters and one-sided error

The Bloom filter [4] is the canonical succinct data structure for approximate set membership. It represents a set $S \subseteq U$ using a bit array of size m and k independent hash functions. A defining property of this structure is its one-sided error guarantee: false negatives never occur—if $u \in S$ the filter always returns 'yes'—but false positives may occur with a controlled probability ε. This asymmetry is critical for many applications: a 'no' answer can always be trusted to gate access to a slower exact lookup. The information-theoretic lower bound for any data structure with false positive rate ε and no false negatives is $n \log(1/\varepsilon)$ bits [5], and Bloom filters achieve this up to a small constant.

Subsequent work produced variants that improve on the basic construction while maintaining one-sided error, achieving the space lower bound exactly with $O(1)$ query time and support for insertions and deletions [6], comparable space with faster practical performance (the cuckoo filter [7]), and reduced construction time (the ribbon filter [8]). All these constructions preserve the one-sided error guarantee, which is what makes them useful in practice.

2.2. Extending Bloom filters to distance-sensitive queries

A natural generalization replaces exact membership with distance-sensitive membership: given query u, determine whether u is close to some element of S under a metric. Kirsch and Mitzenmacher [3] introduced this problem and proposed the first framework using locality-sensitive hash (LSH) families: replacing the independent uniform hash functions in a Bloom filter with LSH functions makes nearby elements collide with higher probability. Their construction allows both false positives and false negatives—the one-sided error guarantee is lost because LSH is not guaranteed to produce collisions for nearby pairs. Hua et al. [9] extended this approach with practical improvements (the locality-sensitive

Bloom filter, LSBF) but false negatives remain unavoidable under this LSH-based paradigm, as LSH schemes do not guarantee collisions for nearby points.

Relative radius, and its removal A subtlety of this construction is that the proximity it resolves is fundamentally relative rather than absolute: the achievable closeness threshold is coupled to the set size n, so one cannot ask the structure to resolve "within Hamming distance r" for a fixed r independent of n—only a fractional window that must widen as n grows. For applications where the meaningful notion of closeness is a fixed number of edits or differing coordinates, this coupling is a genuine obstruction. Goswami et al. [2] removed it by routing distance-sensitivity through a vector signature whose radius parameter r is an explicit, absolute input decoupled from the set size, at the cost of an additive metric-overhead term rather than a coupling to n; this is the non-private baseline our construction builds on.

Arbitrary radius without false negatives Goswami et al. also recovered the Bloom filter's one-sided error guarantee in the distance-sensitive setting for a given radius r. Their key insight is to replace LSH with vector signatures—a succinct variant of CountSketch [10]—that deterministically preserve near pairs ($D(u,s) \leq r$ implies the signatures are close) while probabilistically separating far pairs ($D(u,s) > cr$ implies the signatures differ), achieving tight space bounds of $\Theta(n(r^2/d + \log(1/\varepsilon)))$ bits via a counting argument and Harper's vertex-isoperimetric inequality.

2.3. Privacy of Bloom filters

The privacy properties of Bloom filters have been studied from several angles, with largely negative conclusions: unmodified filters fail to provide reasonable privacy guarantees, an ambiguity that is weak and degrades as the universe size grows [11], a conclusion that extends to counting Bloom filters without modification [12]. These negative results motivate explicitly private filter constructions. Alaggan et al. [13] (BLIP) studied non-interactive differentially private similarity computation on Bloom filters, but focus on privately computing similarities between filters rather than releasing a filter for membership queries, and lack provable utility guarantees. RAPPOR [14] uses Bloom filter encoding combined with randomized response for local DP, but treats the filter as a representational tool rather than the object being protected.

Patel et al. [1] gave the first construction with formal (ε, δ)-DP guarantees and tight utility analysis for the membership problem. Rather than noise injection, they embed S into a random linear system over a finite field, achieving error probability $\alpha = 1/(e^\varepsilon + 1)$ matching their lower bound, while prior noise-injection approaches yield error of order $O(1/\varepsilon)$—exponentially worse. Their space lower bound also applies to probabilistic filters allowing false negatives, establishing the problem's information-theoretic complexity.

2.4. Privacy of distance-sensitive Bloom filters

Though no prior work has analyzed the privacy of the distance-sensitive versions, the two prior distance-sensitive constructions fail differential privacy for complementary reasons, and isolating these reasons clarifies what a differentially private distance-sensitive representation (DPDS-representation) must provide. The first concerns the structure of the error, the second the encoding itself.

Two-sided error is necessary Differential privacy forces an irreducible two-sided error floor on any useful membership, and hence any proximity test. Reducing to the single membership bit, fix neighbors S and $S' = S \cup \{x\}$ with $x \notin S$, and write $a = \Pr[\text{accept } x \mid S']$ (completeness) and $b = \Pr[\text{accept } x \mid S]$ (false-positive rate). Applying (ε, δ)-DP to this binary outcome gives both $a \leq e^\varepsilon b + \delta$ and $1 - b \leq e^\varepsilon(1 - a) + \delta$, hence $b \geq e^{-\varepsilon}(a - \delta)$ and $1 - a \geq e^{-\varepsilon}((1 - b) - \delta)$. For any structure with non-trivial completeness and soundness, both the false-positive rate b and the false-negative rate $1 - a$ are therefore bounded below by $\Omega(e^{-\varepsilon})$ under pure DP; one escapes only by

taking δ large enough to be vacuous. The decisive point is that this error floor is irreducible: privacy forbids driving completeness and soundness simultaneously to one. The vector signature of Goswami et al. [2] has exactly the opposite character. Its only error is the sketch failure probability, which can be made arbitrarily small by enlarging the signature, so the test is faithful and indeed decodable in the limit of more space. A construction whose error tends to zero as more resources are spent cannot exhibit the mandatory two-sided floor, and so cannot be made differentially private; the very faithfulness that makes the signature a good distance estimator is what disqualifies it.

The encoding of neighboring sets must be private A DPDS-representation additionally requires that the released encoding of two sets differing in a single element be indistinguishable: for $|S \triangle S'| = 1$, the distributions over the published structure must satisfy the DP inequalities. The distance-sensitive Bloom filter of Kirsch and Mitzenmacher [3], once its public hash functions are fixed, is a deterministic function of the input set. Two neighboring sets thus map to deterministically distinct filters, and an observer of the structure distinguishes them with certainty—a flat violation of the e^{ε} ratio regardless of how the proximity queries themselves behave. This failure is orthogonal to the first: even a filter with a favorable two-sided query error would remain non-private here, because the leak is in the encoding map $S \mapsto \mathrm{filter}(S)$, not in the answer distribution. The two obstructions are therefore complementary—one on the irreducible error structure demanded of the queries, the other on the indistinguishability demanded of the encoding—and a DPDS-representation must resolve both at once. This is precisely the gap our construction closes by obtaining privacy from the random linear system of Patel et al. [1], whose exclusion sampling supplies the irreducible false negatives, whose randomized solving supplies the false positives, and whose released structure is private under the neighboring-set relation by construction.

2.5. Differentially private linear sketches and sparse histograms

Linear sketches such as CountSketch [10] and Count-Min [15] are a closely related family of succinct structures for frequency estimation and heavy-hitter detection; the fundamental result is that they can be made DP by adding calibrated noise to each entry [16], with per-entry expected error $O(1/\varepsilon)$, which is optimal [17]. Subsequent work has sharpened this for sparse and high-dimensional settings, giving DP sparse vector representations [18] and improved private CountSketch analyses [19, 20], while pan-private variants [21] extend the guarantee to adversaries with access to internal state. These works address the more general problem of frequency estimation, where each element has an associated value; Patel et al. show that restricting to the membership problem (all values are 0 or 1) allows exponentially smaller error—$1/(e^{\varepsilon}+1)$ rather than $\Omega(1/\varepsilon)$, which is impossible for private histograms.

2.6. Positioning of our work

Our work occupies a new position in this landscape. Patel et al. achieve optimal privacy-utility and space tradeoffs for exact membership under DP, but provide no distance-sensitivity. Goswami et al. achieve tight space bounds for distance-sensitive membership without false negatives, but provide no privacy guarantee. Prior DP filter work either lacks provable utility guarantees, applies only to exact membership, or uses Bloom filters as a representational tool for a different problem (e.g., RAPPOR). We combine both frameworks—the Patel et al. linear system for privacy and the Goswami et al. vector signature for distance-sensitivity—to give the first construction achieving both simultaneously for a user-defined radius r. The key insight is that accepting false negatives as part of the error budget α, rather than treating them as a failure, lets us run the Patel et al. mechanism unchanged on the image of S under the Goswami et al. signature and recover distance-sensitivity at query time by expanding each query into an ℓ_1 ball of candidate signatures. Because the signature is fixed public randomness independent of S, the Patel et al. privacy proof carries through unchanged.

3. Preliminaries

3.1. Differential privacy

Definition 3.1 ((ε, δ)-Differential Privacy [16]). *A randomized algorithm M with domain $\mathcal{D}$ is (ε, δ)-differentially private if for all $\mathcal{R} \subseteq \mathrm{Range}(M)$ and all $x, y \in \mathcal{D}$ with $|x - y| = 1$,*

$$\Pr[M(x) \in \mathcal{R}] \leq e^{\varepsilon} \cdot \Pr[M(y) \in \mathcal{R}] + \delta.$$

Throughout, we measure distance between sets S and S' as the symmetric set difference $|S \triangle S'|$, and say S_1, S_2 are neighboring when $|S_1 \triangle S_2| = 1$. For our setting $S \subseteq U$ for some universe U and $S' = S \cup \{t\}$ for some $t \notin S$.

3.2. The Patel et al. framework

Patel et al. [1] represent a set $S \subseteq U$ of size k via a random linear system over a finite field $\mathbb{F}$. Their encoding finds $x \in \mathbb{F}^{\ell}$ satisfying $\mathrm{Row}(s) \cdot x = h(s)$ for all s in a subsampled set $S' \subseteq S$, where $\mathrm{Row} : U \to \mathbb{F}^{1 \times \ell}$ is a random band row function and $h : U \to \mathbb{F}$ is a random hash. Decoding u checks whether $\mathrm{Row}(u) \cdot x = h(u)$.

Theorem 3.2 (Patel et al., Theorems 1.1 and 1.3). *For any $\varepsilon > 0$, $\delta > 0$, there exists an (ε, δ)-DP algorithm representing sets of size k with error probability $\alpha = 1/(e^{\varepsilon} + 1)$ and space $1.05\, k\varepsilon \log e$ bits. Furthermore, any (ε, δ)-DP algorithm for representing sets must have error probability $\alpha \geq (1 - \delta)/(e^{\varepsilon} + 1)$.*

3.3. The Goswami et al. framework

Goswami et al. [2] construct (r, c, ε)-distance-sensitive filters for sets $S \subseteq \{0,1\}^d$: data structures that always return 'yes' for queries within distance r of S, and return 'no' with probability $\geq 1 - \varepsilon$ for queries at distance $> cr$. Their key tool is the vector signature: a map $\sigma : \{0,1\}^d \to \mathbb{Z}^m$ whose ℓ_1 gap $\gamma(x, y) = \|\Gamma(x, y)\|_1$ satisfies $\gamma(x, y) \leq \Psi$ when $D(x, y) \leq r$, and $\gamma(x, y) > \Psi$ with probability $\geq 1 - \varepsilon$ when $D(x, y) > cr$.

4. The DPDS-representation

The two frameworks of Section 3 solve complementary halves of our problem, but they do not compose directly. The Patel et al. linear system provides privacy through a mechanism that is agnostic to what its inputs mean: changing one constraint changes one degree of freedom, and the uniformly random free variable bounds the resulting shift in the output distribution. Its decoder, however, answers only *exact* queries—a query u is checked against the constraint for u itself, so a point merely near an encoded element receives no benefit from that element's constraint. The Goswami et al. signature σ provides exactly the missing distance-sensitivity: it maps Hamming-near pairs to signatures that are provably close in ℓ_1 and Hamming-far pairs to signatures that are far. But "close" is not "equal," and the linear system has no notion of approximate satisfaction.

A natural attempt to bridge this gap is to *quantize* the signature space—to coarsen σ so that provably-close signatures collapse to a common representative and the linear system's exact-equality test sees a match. This does not work. The Goswami et al. guarantee bounds the *aggregate* ℓ_1 gap $\gamma(x, y) = \|\Gamma(x, y)\|_1$ summed over all m coordinates, with a single threshold Ψ separating the near and far cases; it gives no control over any individual coordinate. A per-coordinate quantizer must therefore reconcile two irreconcilable demands: bins wide enough that a near pair (whose deviation may be spread thinly across many coordinates) never straddles a boundary, yet narrow enough that a far pair (whose excess may likewise be spread thinly) crosses one. Because the signature's near and far thresholds coincide at Ψ, converting metric proximity into exact equality is an LSH-type task that provably requires a collision gap, which a gapless signature cannot supply. No choice of bin width satisfies both demands at once.

We instead resolve the mismatch on the *decode* side, leaving the encoding untouched. The encoder is exactly the Patel et al. mechanism applied to the image set $\sigma(S')$ of a subsampled $S' \subseteq S$; since the encoding distribution is unchanged, the Patel et al. privacy proof carries through verbatim, with $\sigma(s)$ playing the role of s. To answer a query u, the decoder enumerates the ℓ_1 ball $Q_u = \{z : \|z - \sigma(u)\|_1 \le \Psi'\}$ of candidate signatures around $\sigma(u)$, where $\Psi' = \Psi/c_{\text{div}}$ is the signature-space radius corresponding to the gap threshold Ψ, and accepts iff some candidate satisfies its linear constraint. The near guarantee places a near witness's signature in Q_u deterministically, so there are no bin-boundary failures; the far guarantee keeps every encoded signature out of Q_u with high probability. Privacy comes for free because σ is fixed public randomness independent of S, and distance-sensitivity comes from the ball expansion at decode time. The cost is concentrated entirely in the number of candidates $N_Q = |Q_u|$, which is $\text{poly}(k/\alpha)$ in the regime $r = O(c \log(k/\alpha))$ and which we account for in Section 5.

As in the exact-membership setting, a key design choice—forced by the impossibility of one-sided error under differential privacy (Section 2.4)—is that we accept false negatives as part of the error probability α, alongside false positives. The remainder of this section makes the construction precise: we first fix the setup and formally define a DPDS-Representation, then construct the signature map and establish its collision properties under ball expansion, and finally assemble the encoding and decoding algorithms.

4.1. Setup

Let $U = \{0,1\}^d$. We are given a set $S \subseteq U$ of size k, distance parameters r and cr with $c \ge 1$, privacy parameters $\varepsilon > 0$ and $\delta \ge 0$, and a target error probability $\alpha < 1/2$.

Definition 4.1 (DPDS-representation)**.** *An algorithm $\Pi = (\Pi.\text{Encode}, \Pi.\text{Decode})$ is an $(r, c, \alpha, \varepsilon, \delta)$-DPDS-Representation for sets of size k from $\{0,1\}^d$ if:*

1. Utility: *For any S with $|S| = k$ and any query set $Q \subseteq U$,*

$$\Pr\left[\exists q \in Q,\ \Pi.\text{Decode}(\hat{S}, q) \neq \mathbf{1}[D(q,S) \le r]\right] \le \alpha|Q|.$$

   Both false positives ($D(u,S) > cr$, decoder returns 1) and false negatives ($D(u,S) \le r$, decoder returns 0) count toward α. Queries with $r < D(q,S) \le cr$ carry no guarantee.
2. Privacy: $\Pi.\text{Encode}$ *is (ε, δ)-DP with respect to neighboring sets.*

4.2. The signature map and decode-side expansion

Raw signature σ Fix a random $m \times d$ integer matrix M where each column has $\delta_{\text{upd}} = O\big(1 + \frac{c}{r}\log\frac{k}{\alpha}\big)$ nonzero entries in $\{-1, +1\}$, placed uniformly at random, with

$$m = O\left(\frac{r}{c-1} + \left(\frac{c}{c-1}\right)^2 \log\frac{k}{\alpha}\right).$$

Let $c_{\text{mod}} = O(c)$, $c_{\text{div}} = O(c)$. Define

$$\sigma(x)_i = \left\lfloor \frac{(Mx)_i \bmod^* c_{\text{mod}}}{c_{\text{div}}} \right\rfloor, \qquad i \in [m],$$

where $\bmod^*$ is the symmetric modulo operator mapping into $[-\lfloor c_{\text{mod}}/2 \rfloor, \lceil c_{\text{mod}}/2 \rceil)$. Each coordinate $\sigma(x)_i$ takes one of at most $R_\sigma := \lceil c_{\text{mod}}/c_{\text{div}} \rceil = O(1)$ values, so the signature lands in a fixed product range $\Sigma \subseteq \mathbb{Z}^m$. The gap vector $\Gamma(x,y)_i = c_{\text{div}}(\sigma(x)_i - \sigma(y)_i \bmod^* c_{\text{mod}})$ and gap $\gamma(x,y) = \|\Gamma(x,y)\|_1$ satisfy, for a threshold $\Psi = O(\delta_{\text{upd}}\, r)$ (Goswami et al., Theorem 4.1):

- $D(x,y) \le r \Rightarrow \gamma(x,y) \le \Psi$ deterministically;
- $D(x,y) > cr \Rightarrow \gamma(x,y) > \Psi$ with probability $\ge 1 - \alpha/k$.

Field embedding Work over $\mathbb{F} = \mathbb{F}_q$ with q a prime power chosen in Theorem 5.2. Because each coordinate of σ has range $R_\sigma = O(1) < q$, the coordinatewise reduction $\sigma(x)_i \bmod q$ is injective on Σ (no wraparound), so we identify $\sigma(x)$ with its image in $\mathbb{F}^m$. The map σ is fixed public randomness, determined by M and generated before and independently of S.

Signature-space ball The gap threshold Ψ is a bound on $\gamma = \|\Gamma\|_1$, and since $\Gamma = c_{\text{div}} \cdot (\sigma(x) - \sigma(y))$ coordinatewise, the corresponding bound in signature space is

$$\Psi' := \Psi / c_{\text{div}} = O\left(\frac{r}{c} + \log \frac{k}{\alpha}\right).$$

We therefore define the candidate set in signature space at radius Ψ':

$$Q_u := \{z \in \Sigma : \|z - \sigma(u)\|_1 \le \Psi'\}.$$

The decoder accepts u iff some $z \in Q_u$ satisfies its linear constraint. The near guarantee places the witness's signature in Q_u; the far guarantee keeps every encoded signature out of Q_u with high probability.

4.3. Encoding and decoding

Encoding is Patel et al. applied to the image set $\sigma(S')$: subsample $S' \subseteq S$ (each element kept with probability $1 - p$), build the linear system with rows $\text{Row}(\sigma(s'))$ and targets $h(\sigma(s'))$ for $s' \in S'$, and solve for x. The encoder is exactly the Patel et al. mechanism run on the signatures $\sigma(S')$ in place of the raw elements, so the encoding distribution—and hence the privacy proof—is untouched. The decoder is the sole locus of distance-sensitivity: it enumerates Q_u and accepts iff any candidate satisfies its constraint.

Algorithm 1 DPDSSet.Decode$(\hat{S}, u)$ — decode-side expansion

if $\hat{S} = (\perp, \perp, \perp, \perp, S)$ **then**
  return $\mathbf{1}[D(u, S) \le r]$
end if
Compute $\sigma(u) \in \mathbb{F}^m$.
Form $Q_u = \{z \in \Sigma : \|z - \sigma(u)\|_1 \le \Psi'\}$ with $\Psi' = \Psi / c_{\text{div}}$.
for all $z \in Q_u$ **do**
  if $\text{Row}(z) \cdot x = h(z)$ **then**
    return 1
  end if
end for
return 0

Lemma 4.2 (Collision properties of σ under ball expansion)**.** *Fix a query u and the encoded image set $\sigma(S')$.*

(i) (Completeness.) *If $D(u, s^*) \le r$ for some $s^* \in S'$, then $\sigma(s^*) \in Q_u$, and the decoder accepts u whenever the constraint for $\sigma(s^*)$ is present and satisfied.*

(ii) (Soundness, image term.) *If $D(u, S) > cr$, then with probability at least $1 - \alpha$ no encoded signature lies in Q_u; that is, $\sigma(s) \notin Q_u$ for all $s \in S$.*

Proof. (i) By the near guarantee, $D(u, s^*) \le r$ gives $\gamma(u, s^*) \le \Psi$ deterministically, hence $\|\sigma(u) - \sigma(s^*)\|_1 = \gamma(u, s^*)/c_{\text{div}} \le \Psi'$, so $\sigma(s^*) \in Q_u$. If $s^* \in S'$ its constraint $\text{Row}(\sigma(s^*)) \cdot x = h(\sigma(s^*))$ holds by construction (assuming Solve succeeded), and the decoder, which tests every $z \in Q_u$, tests

this z and accepts. There is no bin-boundary failure term: the near witness's signature is in the ball deterministically, so the former α_{bin} vanishes.

(ii) Fix $s \in S$ with $D(u,s) > cr$. By the far guarantee, $\gamma(u,s) > \Psi$ with probability $\geq 1 - \alpha/k$, i.e. $\|\sigma(u) - \sigma(s)\|_1 > \Psi'$ and $\sigma(s) \notin Q_u$. A union bound over the $|S| = k$ elements gives $\Pr[\exists s \in S : \sigma(s) \in Q_u] \leq k \cdot (\alpha/k) = \alpha$. □

Remark 4.3. *Lemma 4.2(ii) controls only the event that a* genuine *encoded signature falls into the query ball. It does* not *bound the probability that some non-encoded candidate $z \in Q_u$ spuriously satisfies a constraint; that event is governed by the field size and is handled separately in Theorem 5.2. Separating these two sources is essential—the first is a property of the signature geometry, the second of the random hash h.*

5. Analysis

5.1. Error probability

The decoder's error has three sources: subsampling of the near witness (false negative), a genuine far signature landing in the query ball (false positive, image term, Lemma 4.2(ii)), and a non-encoded candidate in the ball spuriously satisfying a constraint (false positive, hash term). We split the false-positive budget evenly between the image and hash terms: the signature failure probability is set to $\alpha/(2k)$ and the field size to absorb the N_Q-fold union over candidates.

Lemma 5.1 (Candidate count)**.** *Let $R_\sigma = \lceil c_{\text{mod}}/c_{\text{div}} \rceil = O(1)$ and $\Psi' = \Psi/c_{\text{div}} = O\big(\frac{r}{c} + \log \frac{k}{\alpha}\big)$. Then*

$$|Q_u| \leq \sum_{j=0}^{\Psi'} \binom{m}{j} (2R_\sigma)^j = 2^{\,\Theta(\Psi' \log(R_\sigma\, m/\Psi'))},$$

so with $m = O\big(\frac{r}{c-1} + (\frac{c}{c-1})^2 \log \frac{k}{\alpha}\big)$ one has $\ln |Q_u| = \Theta\big(\Psi' \log(m/\Psi')\big)$. Explicitly, using $\binom{m}{j} \leq (em/j)^j$ termwise,

$$|Q_u| \;\leq\; \left(\frac{2e\, R_\sigma\, m}{\Psi'}\right)^{\Psi'},$$

and in the regime $r = O\big(c \log(k/\alpha)\big)$ with c bounded away from 1 one has $\Psi' = O(\log(k/\alpha))$ and $m/\Psi' = O(c/(c-1)) = O(1)$, so

$$|Q_u| \;\leq\; \left(\frac{k}{\alpha}\right)^{D(c)}, \qquad D(c) = \frac{\Psi'}{\log(k/\alpha)} \cdot \log_2 \frac{2e\, R_\sigma\, m}{\Psi'},$$

a constant determined by the closed form above. The exponent does not collapse to 1 as c grows: even in the large-c limit $m/\Psi' \to 1^+$ leaves $D(c) = \Theta(1)$ (numerically $D(c) \approx 5$ for $c = 16$), and $D(c)$ increases as $c \to 1^+$ through the $(c/(c-1))^2$ factor in m (numerically $D(c) \approx 7$ for $c = 2$). Hence $\ln |Q_u| = O(\log(k/\alpha))$.

Proof. A point $z \in \Sigma$ with $\|z - \sigma(u)\|_1 \leq \Psi'$ differs from $\sigma(u)$ in at most Ψ' coordinates, each contributing at least 1 to the ℓ_1 distance. Choosing the $j \leq \Psi'$ differing coordinates in $\binom{m}{j}$ ways and assigning each a nonzero signed offset within the range R_σ in at most $2R_\sigma$ ways gives $|Q_u| \leq \sum_{j=0}^{\Psi'} \binom{m}{j} (2R_\sigma)^j$. Using $\binom{m}{j} \leq (em/j)^j$ and summing the geometric-like series, the total is at most $(2eR_\sigma m/\Psi')^{\Psi'}$, the largest term up to a poly factor. In the stated regime both Ψ' and m are $O(\log(k/\alpha))$ with $m/\Psi' = O(c/(c-1))$, so taking $\log_2$ and dividing by $\log(k/\alpha)$ gives the exponent $D(c)$ above; it is $\Theta(1)$ for c bounded away from 1 and grows as $c \to 1^+$. Hence $\ln |Q_u| = O(\log(k/\alpha))$. □

Theorem 5.2 (Error probability)**.** *Suppose $r = O\big(c \log(k/\alpha)\big)$ with c bounded away from 1, so by Lemma 5.1 $|Q_u| \leq N_Q := (k/\alpha)^{D(c)}$ with $D(c) = \Theta(1)$ the closed-form exponent of Lemma 5.1. Set*

the signature failure probability to $\alpha/(2k)$, the field size to $q = \lceil 2N_Q/\alpha \rceil$ (a prime power), and the subsampling probability to $p = 1/(q-1)$. Then DPDSSet *has error probability at most α for both false positives and false negatives, independently across queries.*

Proof. False positives. Suppose $D(u, S) > cr$. By Lemma 4.2(ii) with signature failure $\alpha/(2k)$, a union bound over $|S| = k$ gives $\Pr[\exists s \in S : \sigma(s) \in Q_u] \leq k \cdot \frac{\alpha}{2k} = \frac{\alpha}{2}$ (image term). Condition on its complement. Then every candidate $z \in Q_u$ has its constraint absent from the system, so $h(z)$ is independent uniform and $\Pr[\mathrm{Row}(z) \cdot x = h(z)] = q^{-1}$. A union bound over $|Q_u| \leq N_Q$ gives $N_Q/q \leq \alpha/2$ (hash term). Total false positive $\leq \alpha$.

False negatives. Suppose $D(u, S) \leq r$ with witness s^*. If $s^* \in S'$, then by Lemma 4.2(i) $\sigma(s^*) \in Q_u$ deterministically and its constraint is satisfied, so the decoder accepts; there is no bin-boundary failure term. The only failure is $s^* \notin S'$, of probability $p = 1/(q-1) \leq \alpha$. Hence false negative $\leq \alpha$.

Independence. For distinct queries contributing distinct non-encoded signatures, the corresponding h-values are independent uniform, giving independent error events; queries whose balls share candidates share those h-values, exactly as the shared-element caveat in Patel et al. □

Remark 5.3 (The regime, and the $c \to 1$ limitation)**.** *The condition $r = O\big(c \log(k/\alpha)\big)$ is the boundary of usefulness, and it widens with the approximation factor: a larger c coarsens the signature (smaller m and Ψ' relative to r), admitting a proportionally larger absolute radius. Within the regime the additive privacy penalty over exact membership is $\Delta\varepsilon = \ln N_Q = \Theta(\log(k/\alpha))$; this is a constant-factor blowup over Patel's $\varepsilon \approx \log(1/\alpha)$ (constant when $k = \mathrm{poly}(1/\alpha)$), and it does* not *vanish as $r \to 0$—the signature length alone forces $\Psi' = \Theta(\log(k/\alpha))$. The penalty constant degrades as $c \to 1^+$: the $(c/(c-1))^2$ factor in m inflates m/Ψ', so the construction is most effective for c bounded away from 1 and approaches the naive Bloom-filter ball-query baseline in the exact-membership limit. The factor c remains a free user parameter throughout.*

5.2. Differential privacy

Theorem 5.4 (Differential privacy)**.** *With field size $q = \lceil 2N_Q/\alpha \rceil$ and subsampling probability $p = 1/(q-1)$,* DPDSSet.Encode *is (ε, δ)-DP with $\varepsilon = \ln(q-1)$ and $\delta = f_{\mathrm{Solve}} \leq 2^{-\Omega(w)}$, where $w = O(\log(1/\delta) + \log k)$ is the band width.*

Proof. The proof is identical to Patel et al. (Theorem B.1), with $\sigma(s)$ playing the role of s throughout. The choice $p = 1/(q-1)$ is exactly the Patel et al. subsampling probability $\alpha_0/(1-\alpha_0)$ expressed for a field of size q (their $\alpha_0 = 1/q$), so the two DP inequalities follow verbatim: conditioned on excluding the differing element, the two systems are identically distributed, giving one direction with factor $1/p = q - 1$; conditioned on retaining it, the single extra row introduces one free variable set uniformly in $\mathbb{F}$, giving the other direction. Combining yields $\varepsilon = \ln(q-1)$.

The only change from the exact-membership setting is the *value* of q: decode-side ball expansion enlarges the field from $q \approx 1/\alpha$ to $q = \lceil 2N_Q/\alpha \rceil$, so $\varepsilon = \ln(q-1) = \log \frac{1}{\alpha} + \Theta(\log \frac{k}{\alpha})$ rather than $\ln((1-\alpha)/\alpha)$. The privacy argument is unchanged because σ is fixed public randomness determined by M, generated before and independently of S; the argument is therefore entirely insensitive to the metric structure encoded by σ. □

5.3. Space and complexity

The encoding $\hat{S} = (x, \mathrm{Row}, h, \sigma)$ consists of the linear system solution x, the hash functions (Row, h), and the signature map σ (represented by the matrix M). Relative to exact membership, only the field size changes: where exact membership uses $q \approx 1/\alpha$, decode-side expansion requires $q = \lceil 2N_Q/\alpha \rceil$ with $N_Q = \mathrm{poly}(k/\alpha)$ (Theorem 5.2), so each field element now costs $\log q = \log \frac{1}{\alpha} + \log 2N_Q$ bits rather than $\log(1/\alpha)$. This one change accounts for every term below.

It is most visible in the solution vector $x \in \mathbb{F}^\ell$, the dominant k-dependent cost, where $\ell = (1+\beta)k'$ and $\mathbb{E}[k'] = (1-p)k$:

$$\mathbb{E}[|x|] = (1+\beta)\,k(1-p)\big(\log\tfrac{1}{\alpha} + \log 2N_Q\big) = \underbrace{(1+\beta)\,k\,\varepsilon_0 \log e}_{\text{Patel term}} + O\big(k \log N_Q\big),$$

where $\varepsilon_0 = \ln(1/\alpha)$. The first term is exactly the Patel et al. cost; the second, the enlarged-field overhead, is $O(k\log(k/\alpha))$ bits in the regime $r = O(c\log(k/\alpha))$. The linear system still sees $|\sigma(S')| = k'$ constraints, so only the *width* of each field element grows, not their count.

The metric structure, by contrast, is untouched by the larger field: storing M's δ_{upd} nonzero ± 1 positions per column, at $O(\log m)$ bits each, still costs $O\big(d\,\delta_{\text{upd}} \log m\big) = O\big(d(1 + \frac{c}{r}\log\frac{k}{\alpha})\log m\big)$ bits, independent of k except logarithmically—now the *only* cost of the metric structure. The hash functions add the usual lower-order terms, $O(w \log q)$ for the band Row and $O(\log q)$ for h, $w = O(\log(1/\delta) + \log k)$. Together, the total space is

$$|\hat{S}| = (1+\beta)\,k\,\varepsilon_0 \log e + O\big(k \log \tfrac{k}{\alpha}\big) + O\Big(d\Big(1 + \tfrac{c}{r}\log\tfrac{k}{\alpha}\Big)\log m\Big) \text{ bits},$$

a clean three-way split: the Patel reference cost, the space-side image of the privacy penalty $\Delta\varepsilon = \log N_Q$, and the k-independent metric overhead.

The running time follows the same logic. Encoding evaluates $\sigma(s)$ for each of the k' subsampled elements ($O(d\,\delta_{\text{upd}})$ each) and solves the band system in $O(k'w)$, giving $O\big(kd(1+\frac{c}{r}\log\frac{1}{\alpha}) + k(\log\frac{1}{\delta} + \log k)\big)$ —unchanged from exact membership up to arithmetic over the larger field. Decoding is where the repair costs something algorithmically: rather than the single dot product exact membership performs, the decoder evaluates $\sigma(u)$ once and tests every candidate in Q_u, giving

$$O\Big(d\Big(1 + \tfrac{c}{r}\log\tfrac{1}{\alpha}\Big) + N_Q \cdot w\Big)$$

—an N_Q-fold blowup, $\text{poly}(k/\alpha)$ precisely in the regime $r = O(c\log(k/\alpha))$, and the one place in this accounting where the construction pays in a currency exact membership does not.

5.4. Main result

The discussion and analysis above give us our main result which we present here.

Theorem 5.5 (Main result). *Let $S \subseteq \{0,1\}^d$ with $|S| = k$, and let $r = O\big(c\log(k/\alpha)\big)$ with approximation factor c bounded away from 1. Set $q = \lceil 2N_Q/\alpha \rceil$ with $N_Q = (k/\alpha)^{D(c)}$ and $D(c) = \Theta(1)$ the closed-form exponent of Lemma 5.1, subsampling $p = 1/(q-1)$, and signature failure $\alpha/(2k)$. Then* DPDSSet *is an $(r, c, \alpha, \varepsilon, \delta)$-DPDS-Representation with:*

- *(i) $\varepsilon = \ln(q-1) = \ln\frac{1}{\alpha} + \ln 2N_Q$, where $\ln N_Q = \Theta\big(\Psi' \log(m/\Psi')\big) = \Theta\big(\log\frac{k}{\alpha}\big)$ in the stated regime. The second term dominates: $\ln 2N_Q$ exceeds $\ln(1/\alpha)$ by a c-dependent factor (a small constant, growing as $c \to 1^+$ and of order 4–8 in typical regimes), so $\varepsilon = \Theta\big(\log\frac{k}{\alpha}\big)$ is governed by the expansion penalty, not by the exact-membership value $\ln(1/\alpha)$. Also $\delta \le 2^{-\Omega(w)}$;*
- *(ii) error $\le \alpha$ for both false positives and false negatives, independently across queries;*
- *(iii) space $(1+\beta)k\,\varepsilon_0 \log e + \Theta\big(k \log N_Q\big) + O\big(d(1+\frac{c}{r}\log\frac{k}{\alpha})\log m\big)$ bits, where $\varepsilon_0 = \ln(1/\alpha)$ is the exact-membership reference value (*not *this construction's privacy parameter ε); the second term $\Theta(k\log N_Q) = \Theta(k\log\frac{k}{\alpha})$ is the enlarged-field overhead and, as in (i), dominates the first by the same c-dependent factor;*
- *(iv) encoding time $O\big(kd(1+\frac{c}{r}\log\frac{1}{\alpha}) + k(\log\frac{1}{\delta} + \log k)\big)$;*
- *(v) decoding time $O\big(d(1+\frac{c}{r}\log\frac{1}{\alpha}) + N_Q \cdot w\big)$, $w = O(\log\frac{1}{\delta} + \log k)$.*

Comparison with prior work The privacy parameter retains the exact-membership functional form of Patel et al., but the enlarged field makes it $\Theta(\log(k/\alpha))$-factor worse, not a lower-order correction (part (i)); we do *not* claim optimality, since the Patel et al. lower bound is for exact membership only and whether this penalty is necessary for distance-sensitive queries is open (Section 6). The space, similarly dominated by the enlarged-field overhead, is otherwise notable for its k-independent metric term: the Goswami et al. non-private filter stores one signature per element and pays the metric cost k times over, while our single shared matrix M does not. The approximation factor c is free throughout; the construction is most effective for c bounded away from 1 and $r = O(c\log(k/\alpha))$, degrading toward a naive ball query outside this regime or as $c \to 1^+$.

6. Future work

Our construction opens several directions that we plan to pursue. We describe each below, together with our intended approach and the technical obstacles we expect to encounter, organized roughly by tractability.

6.1. Unified space lower bound

The most immediate next step is establishing that the space usage of our construction is necessary. Specifically, we aim to prove a conjecture that any $(r, c, \alpha, \varepsilon, \delta)$-DPDS-Representation Π for sets of size k from $\{0,1\}^d$ with $0 < \alpha \le 1/2$ must satisfy

$$\mathbb{E}[s] \ge \Omega\left(k\varepsilon + k\left(\frac{r^2}{d} + \log\frac{1}{\alpha} - H(\alpha)\right)\right),$$

where $H(\alpha) = -\alpha\log\alpha - (1-\alpha)\log(1-\alpha)$ is the binary entropy function.

Starting point The two existing lower bound arguments provide the ingredients. The Patel et al. space lower bound uses a compression argument: a (ε, δ)-DP encoding with error α implies an efficient compression of sparse binary vectors, contradicting Shannon's theorem. The Goswami et al. lower bound uses a counting argument: the positive set $P_S = \{u : \mathrm{Decode}(\hat{S}, u) = 1\}$ must be large enough to contain all of S (using the no-false-negatives assumption), and the vertex-isoperimetric inequality (Harper's theorem) bounds how large the r-erosion P_S^{-r} can be. We have already verified that the Goswami et al. argument can be repaired to handle false negatives at an additive cost of $kH(\alpha)$ bits, giving a corrected bound of $\Omega(k(\log(1/\alpha) - H(\alpha)))$ for the counting term, while the $\Omega(kr^2/d)$ isoperimetric term is unaffected (it only requires an upper bound on $|P_S^{-r}|$, which becomes smaller when false negatives shrink P_S).

Planned approach We plan to construct a unified compression scheme that decomposes $S = (S \cap P_S) \cup (S \setminus P_S)$, encodes the false negatives $S \setminus P_S$ at cost $kH(\alpha)$ bits, encodes the true positives $S \cap P_S$ as a subset of P_S^{-r} at cost $\log\binom{|P_S^{-r}|}{|S\cap P_S|}$ bits, and applies Shannon's source coding theorem using the DP constraint to lower-bound the conditional entropy $H(S \mid \hat{S})$.

Expected obstacles Three incompatibilities between the existing arguments must be resolved: the no-false-negatives assumption in the Goswami et al. isoperimetric step, where the containment $B(s, r, d) \subseteq P_S$ fails for $S \setminus P_S$ even after our counting-term repair; the DP constraint being distributional while the counting argument is combinatorial on the family $\{P_S\}$; and the two compression schemes' differing structures (Patel et al. partitions the universe into decoded-0/decoded-1 sets, Goswami et al. encodes S as a subset of P_S). The central obstacle, where we expect the main effort to lie, is converting the (ε, δ)-DP constraint into a quantitative lower bound on $H(S \mid \hat{S})$: the pure-DP bound $I(S; \hat{S}) \le \varepsilon\,\mathbb{E}[|S|]\log e$ is too loose to recover the Patel et al. term with the right constants, and tightening this conversion—or finding an alternative that avoids it—is the core of the problem.

6.2. Privacy-utility lower bound

Our construction achieves privacy parameter $\varepsilon = \ln(q-1) = \ln(1/\alpha) + \Theta(\log(k/\alpha))$, a $\Theta(\log(k/\alpha))$-factor above the Patel et al. exact-membership tradeoff $\ln((1-\alpha)/\alpha)$ that the decode-side expansion forces (Theorem 5.5). The Patel et al. lower bound is proven only for exact membership and does not apply to the distance-sensitive setting, so we cannot conclude that our penalty is necessary. We intend to investigate whether an analogous bound $\alpha \geq g(\varepsilon, \delta, r, c, d)$ holds for DPDS-Representations, and in particular whether g depends on the metric parameters r, c, d at all, or reduces to the Patel et al. bound $(1-\delta)/(e^{\varepsilon}+1)$. Of special interest is whether the approximation gap $c > 1$ permits strictly better tradeoffs than the exact case $c = 1$: a separation here would show that distance-sensitivity genuinely changes the privacy-utility landscape, while a matching bound would establish optimality of the leading term of our construction. A particularly pointed question is whether the $\Theta(\log(k/\alpha))$ expansion penalty in our ε is intrinsic to converting metric proximity into exact-equality matching, or an artifact of the decode-side approach.

6.3. Randomness complexity

Canonne et al. [22] initiate the study of randomness complexity for DP mechanisms, showing $\log_2 d \pm O(1)$ random bits suffice and are necessary for d-dimensional DP summation, via a connection to deterministic rounding schemes. We see this as a two-stage program. First, the randomness complexity of exact membership: the Patel et al. pure-DP Vandermonde construction uses $O(k + \log(1/\delta))$ random bits, and we ask whether $O(\log k)$ suffices, as in the summation setting. This looks approachable, since both the upper-bound technique (rounding schemes via output entropy) and the lower-bound technique (extracting a rounding scheme from an accurate, low-randomness mechanism) are general frameworks that plausibly need only the right combinatorial analog of a rounding scheme for sets, not a new theorem. Second, we extend the analysis to distance-sensitive representations, where randomness also enters through the signature matrix M; our construction uses $O(d \cdot \delta_{\text{upd}} + k + w\log(1/\alpha))$ random bits, and we will investigate whether this is optimal.

6.4. Extensions to other metrics

Extending beyond Hamming space requires two ingredients, easy to conflate: a signature σ with the collision property of Lemma 4.2 (near pairs collide, far pairs separate under ball expansion), and a *ball-counting argument* analogous to Lemma 5.1 bounding $N_Q = \text{poly}(k/\alpha)$, without which Theorems 5.2 and 5.5 give no useful tradeoff regardless of the collision property. For ℓ_1 balls over the bounded-range lattice Σ this count is a standard combinatorial bound; it is not inherited for free elsewhere—in particular, ℓ_2 ball volumes grow with dimension differently than ℓ_1, so our regime $r = O(c\log(k/\alpha))$ need not carry over even with a suitable ℓ_2 signature. We will pursue ℓ_2 first, given existing private linear sketch machinery [19], then Hamming distance over larger alphabets and the Hausdorff metric on point sets. Edit distance remains challenging even non-privately due to the lack of a suitable LSH family [3]; whether our construction offers leverage there, on either ingredient, is left for later.

7. Conclusion

We have presented the first (ε, δ)-DP distance-sensitive set representation for Hamming space. Our construction runs the Patel et al. linear system framework unchanged on the image of S under the Goswami et al. vector signature σ, and recovers distance-sensitivity on the decode side by expanding each query into the ℓ_1 ball of candidate signatures. The privacy proof carries through unchanged because σ is fixed public randomness independent of S, and accepting false negatives as part of the error budget α keeps the two frameworks compatible. In the regime $r = O(c\log(k/\alpha))$ the construction achieves error at most α on both sides (Theorem 5.5), at the cost of a $\Theta(\log(k/\alpha))$-factor penalty over exact membership in both the privacy parameter and the dominant space term; the approximation

factor c remains free throughout. Several open problems remain, of which the unified space lower bound and the privacy-utility lower bound for the distance-sensitive setting are the most important for establishing the optimality of our result.

Declaration on generative AI

During the preparation of this work, the author used Claude and overleaf in order to do minor grammar revision, spelling check, paraphrase and reword. After using these tools, the author reviewed and edited the content as needed and assumes full responsibility for the content of the publication.

References

[1] S. Patel, G. Persiano, J. Y. Seo, K. Yeo, Differentially private set representations, Advances in Neural Information Processing Systems 37 (2024) 81888–81911. doi:10.52202/079017-2602.

[2] M. Goswami, R. Pagh, F. Silvestri, J. Sivertsen, Distance sensitive bloom filters without false negatives, in: Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2016. doi:10.1137/1.9781611974782.17.

[3] A. Kirsch, M. Mitzenmacher, Distance-sensitive bloom filters, in: Proceedings of the 8th Workshop on Algorithm Engineering and Experiments (ALENEX), 2006, pp. 41–50.

[4] B. H. Bloom, Space/time trade-offs in hash coding with allowable errors, Communications of the ACM 13 (1970) 422–426. doi:10.1145/362686.362692.

[5] A. Broder, M. Mitzenmacher, Network applications of bloom filters: A survey, Internet Mathematics 1 (2004) 485–509.

[6] R. Pagh, A. Pagh, S. S. Rao, An optimal bloom filter replacement, in: Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2005, pp. 823–829.

[7] B. Fan, D. G. Andersen, M. Kaminsky, M. D. Mitzenmacher, Cuckoo filter: Practically better than bloom, in: Proceedings of the 10th ACM International Conference on Emerging Networking Experiments and Technologies (CoNEXT), 2014, pp. 75–88. doi:10.1145/2674005.2674994.

[8] P. C. Dillinger, S. Walzer, Ribbon filter: Practically smaller than Bloom and Xor, CoRR abs/2103.02515 (2021). doi:10.4230/LIPIcs.SEA.2022.4.

[9] B. Hua, Y. Xiao, B. Veeravalli, D. Feng, Locality-sensitive bloom filter for approximate membership query, IEEE Transactions on Computers 61 (2012) 817–830. doi:10.1109/TC.2011.108.

[10] M. Charikar, K. Chen, M. Farach-Colton, Finding frequent items in data streams, in: Automata, Languages and Programming (ICALP 2002), 2002, pp. 693–703.

[11] G. Bianchi, L. Bracciale, P. Loreti, "Better Than Nothing" privacy with bloom filters: To what extent?, in: Privacy in Statistical Databases (PSD 2012), volume 7556 of *Lecture Notes in Computer Science*, Springer, 2012, pp. 348–363. doi:10.1007/978-3-642-33627-0_27.

[12] P. Reviriego, A. Sánchez-Macian, S. Walzer, E. Merino-Gómez, S. Liu, F. Lombardi, On the privacy of counting bloom filters, IEEE Transactions on Dependable and Secure Computing 20 (2022) 1488–1499. doi:10.1109/TDSC.2022.3158469.

[13] M. Alaggan, S. Gambs, A.-M. Kermarrec, BLIP: Non-interactive differentially-private similarity computation on bloom filters, in: Stabilization, Safety, and Security of Distributed Systems (SSS 2012), Springer, 2012, pp. 202–216.

[14] Ú. Erlingsson, V. Pihur, A. Korolova, RAPPOR: Randomized aggregatable privacy-preserving ordinal response, in: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS), 2014, pp. 1054–1067. doi:10.1145/2660267.2660348.

[15] G. Cormode, S. Muthukrishnan, An improved data stream summary: The count-min sketch and its applications, Journal of Algorithms 55 (2005) 58–75. doi:10.1016/j.jalgor.2003.12.001.

[16] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of Cryptography Conference (TCC 2006), 2006, pp. 265–284. doi:10.1007/11681878_14.

[17] M. Hardt, K. Talwar, On the geometry of differential privacy, in: Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC), 2010, pp. 705–714.

[18] M. Aumüller, C. J. Lebeda, R. Pagh, Differentially private sparse vectors with low error, optimal space, and fast access, in: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS), 2021, pp. 1223–1236.

[19] F. Zhao, D. Qiao, R. Redberg, D. Agrawal, A. E. Abbadi, Y.-X. Wang, Differentially private linear sketches: Efficient implementations and applications, in: Advances in Neural Information Processing Systems (NeurIPS 2022), volume 35, 2022, pp. 12691–12704. doi:10.5555/3600270.3601192.

[20] R. Pagh, M. Thorup, Improved utility analysis of private CountSketch, in: Advances in Neural Information Processing Systems (NeurIPS 2022), volume 35, 2022, pp. 25631–25643. doi:https://doi.org/10.52202/068431-1858.

[21] D. J. Mir, S. Muthukrishnan, A. Nikolov, R. N. Wright, Pan-private algorithms via statistics on sketches, in: Proceedings of the 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), 2011, pp. 37–48. doi:10.1145/1989284.1989290.

[22] C. L. Canonne, F. E. Su, S. P. Vadhan, The randomness complexity of differential privacy, in: 16th Innovations in Theoretical Computer Science Conference (ITCS 2025), volume 27 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 2025, pp. 27:1–27:21. doi:10.4230/LIPIcs.ITCS.2025.27.