Lightweight Private Similarity Search

Sacha Servan-Schreiber
MIT CSAIL
3s@mit.edu

Simon Langowski
MIT CSAIL
slangows@mit.edu

Srinivas Devadas
MIT CSAIL
devadas@mit.edu

Nearest neighbor search is a fundamental building-block for a wide range of applications. A privacy-preserving protocol for nearest neighbor search involves a set of clients who send queries to a remote database. Each client retrieves the nearest neighbor(s) to its query in the database *without* revealing any information about the query. For database privacy, the client must not learn anything beyond the query answer.

Existing protocols for private nearest neighbor search require heavy cryptographic tools, resulting in poor practical performance or large client overheads. In this paper, we present the first *lightweight* protocol for private nearest neighbor search. Our protocol is instantiated using two non-colluding servers, each holding a replica of the database. The protocol supports an arbitrary number of clients simultaneously querying the database via these servers. Each query is only a single round of communication for the client and does not require any communication between servers.

If at least one of the servers is non-colluding, we ensure that (1) no information is revealed on the client's query, (2) the total communication between the client and the servers is sublinear in the database size, and (3) each query answer only leaks a small and precisely quantified amount of information about the database to the client, even when the client is acting maliciously.

We implement our protocol and report its performance on real-world data. Our construction requires between 10 and 30 seconds of query latency over large databases of 10M feature vectors. Client overhead remained under 10 µs of processing time per query and typically less than 4 MB of communication, depending on parameters.

1 Introduction

Nearest neighbor search is used in a wide range of online applications, including recommendation engines [34, 111], reverse image search [74], image-recognition [77], earthquake detection [118], computational linguistics [84], natural-language processing [99], targeted advertising [97, 112], and numerous other areas [71, 78, 91, 97].

In these settings, a server has a database of high-dimensional feature vectors associated with items. Clients send query vectors to the server to obtain the set of items (a.k.a. neighbors) that have similar vectors relative to the issued query. Typically, the client only obtains the identifiers (IDs) of the neighbors rather than the feature vectors themselves, as the server may wish to keep the feature vectors private. The IDs of the feature vectors can be documents, songs, or webpages, and therefore all the client requires as output for correct functionality.

For a concrete example, consider a music recommendation engine such as Spotify. The Spotify server holds a database of song feature vectors. Each feature vector can be seen as a concise representation of song attributes — e.g., genre, popularity, user ratings — encoded in a high-dimensional vector space. A Spotify user has a vector of features (the query) representing their musical interests. The goal is to recommend songs the user may find interesting, which should have similar features. This is done using nearest neighbor search to find the ID (e.g., the song) of a vector similar to the query. The client learns the recommended song *without* learning the potentially proprietary feature vector associated with it (beyond what can be implicitly inferred through similarity with the query).

In the above example, the Spotify database learns exactly which music genres (singers, etc.) the client is interested in. It is not difficult to see that in applications that involve more sensitive user data, similarity search can easily violate user privacy. Such applications include targeted advertising [21, 61, 98, 109, 112], biometric data [23, 49], medical records [15, 104], and DNA data analysis [35, 81, 89, 115]. These applications construct queries from highly personal user information. For example, in the medical setting, a person's medical history, demographics, and even DNA can be compiled into a query. The resulting neighbors are other people who have similar symptoms, gene sequences, or health conditions [87]. Both regulatory and ethical reasons dictate that such personal information (represented in the query) should be kept private from the database.

The potential privacy issues surrounding similarity search have motivated a handful of privacy-preserving protocols [36,

65, 95, 104, 119]. However, due to the complexity of the problem, existing protocols are highly inefficient. Prior work either makes use of *heavy* cryptographic tools (e.g., two-party computation and fully-homomorphic encryption) or fails to provide strong privacy guarantees for users (e.g., leak information on the query to the server). See the overview of related work in Section 9.

Additionally, existing protocols do not consider *malicious*[1] clients that may attempt to abuse the system to learn more information about the database. The proposed solutions leave open the problem of designing a concretely efficient protocol for private nearest neighbor search, especially when strong security guarantees are required.

The contribution of this paper is the design and implementation of a *lightweight* protocol for private similarity search. Specifically, we provide a private protocol for solving the Approximate Nearest Neighbor (ANN) search problem. Our protocol provides strong security guarantees for both the client and the database. Moreover, our protocol is concretely efficient and requires little communication between the client and the database servers (and no communication between servers). We achieve this without compromising on privacy for the client — nothing is leaked on the client's query. For database privacy, our construction requires some extra database leakage compared to the *ideal leakage* which only reveals the ANN to the client. The additional leakage is minimal and allows us to eschew oblivious comparisons, which are highly inefficient to instantiate [116]. However, we are careful to precisely quantify this extra leakage relative to the ideal functionality. In our analysis, we show that the leakage is at most a constant factor worse than the ideal leakage (asymptotically optimal). Moreover, we ensure that malicious clients cannot abuse the protocol to learn more than an honest client would have.

Private ANN search. We operate in the following model (see Figure 1 for a simplified illustration). Fix a database $\mathcal{DB}$ containing a set of N feature vectors $\boldsymbol{v_1}, \ldots, \boldsymbol{v_N}$ and corresponding IDs $\mathsf{ID}_1, \ldots, \mathsf{ID}_N$. A client has a query vector $\boldsymbol{q}$. The client must learn only the ID(s) of the nearest neighbor(s) relative to $\boldsymbol{q}$ under some similarity (or distance) metric. This is the standard setup considered in prior work [36, 65, 95, 104, 119]. For client privacy, the protocol must not leak any information about $\boldsymbol{q}$ to the database servers. For database privacy, the protocol must leak as little as possible on $\boldsymbol{v_1}, \ldots, \boldsymbol{v_N}$ to the client. Observe that perfect database privacy (i.e., no database leakage) is unattainable because the client *must* learn at least one ID corresponding to a neighboring vector in the database, which indicates that the vector associated with the ID is similar to the query. We therefore focus on minimizing extra leakage of the database to the client. That is, leaking as little as possible beyond what can be implicitly inferred from the ID of the nearest neighbor.

[1]Existing work requires secure function evaluation between the client and server, which can be "upgraded" to malicious-security at the cost of computationally expensive transformations based on zero-knowledge proofs [57]. The use of fully-homomorphic encryption is another alternative to provide malicious security but comes at a high efficiency cost.

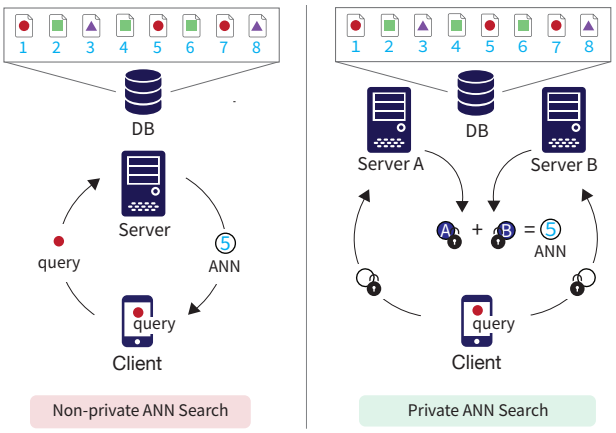

Figure 1: Overview of approximate nearest neighbor search and the privacy-preserving protocol considered in this paper. In the private setting, a client with a query (red dot) interacts with a remote database via two non-colluding servers (Servers A and B). The client combines the responses from both servers to obtain the approximate nearest neighbor ID (in this case the ANN ID = 5) without revealing the query to the servers.

Challenges. Due to the nature of the ANN search problem, privacy-focused solutions use some form of (oblivious) distance comparisons between the query and the feature vectors in the database. Oblivious comparisons [116] require the use of heavy cryptographic tools (e.g., two-party computation or fully-homomorphic encryption). Two-party computation introduces heavy communication costs (typically 1-6 GB per ANN query [36]). Solutions based on fully-homomorphic encryption introduce high computational overhead for the server (hours or even days of processing time [104, 119]), resulting in significant response latency. In this work, we begin by asking the following question:

> *Is it possible to design a private similarity search protocol that is concretely efficient for both the client and the servers while preserving query privacy* ***and*** *database privacy?*

We answer this question in the affirmative.

Our approach. We begin by making several observations about the ANN search problem from both an algorithmic and privacy-preserving perspective. We redesign the standard locality-sensitive-hashing based data structure for ANN search with the goal of avoiding oblivious comparisons (the efficiency bottleneck of prior work). We achieve this by replacing brute-force comparisons with a radix-sort [72] inspired approach for extracting the nearest neighbor. We then show how to query this new data structure through a novel privacy-preserving protocol. We use distributed point functions [54] as an existing building block to querying for neighbors. We apply an efficient leakage-suppressing transformation to query answers (introduced in

Section 5.2) that prevents malicious clients from learning more than one ID per query. Finally, we optimize our protocol by applying partial batch retrieval (a new spin on batch-PIR [14]), which we introduce in Section 6.2.

We show that our protocol is (1) *accurate* through an extensive theoretical analysis and empirical evaluation, (2) *private* by analyzing the security properties of our protocol with respect to client and database privacy, and (3) *efficient* in terms of concrete server-side computation and client-server communication. See Sections 7 and 8 for analytical and empirical results.

Contributions. In summary, this paper makes the following four contributions:

1. design of a single-round protocol for privacy-preserving ANN search, achieving sublinear communication and concrete efficiency,
2. leakage analysis with quantifiable database privacy, which we show asymptotically matches the ideal functionality,
3. security against malicious clients that may deviate from protocol in an attempt to abuse leakage, and
4. an open-source implementation [2] which we evaluate on real-world data with millions of feature vectors.

Limitations. Our protocol has greater database leakage compared to the ideal functionality. We show that the database leakage is asymptotically optimal, but concretely a small factor worse on real data; see empirical analysis in Section 7. Additionally, in contrast to prior work, our threat model assumes two non-colluding servers. We note, however, that using non-colluding servers to instantiate lightweight privacy-preserving systems has proven fruitful in other areas [5, 27, 37, 42, 43, 45, 46, 48, 73, 88], including systems deployed in industry [45, 59]. In our protocol, the servers do not communicate with one another.

2 Background: nearest neighbor search

We begin by describing the standard (non-private) approach to approximate nearest neighbor search based on locality-sensitive hashing. Even outside of a privacy-preserving context, nearest neighbor search in higher dimensions ($d \geq 10$) requires tolerating approximate results to achieve efficient solutions [83]. In Section 4, we transform the ideas from non-private ANN search into a private protocol instantiated between a client and two servers holding replicas of the database.

2.1 Locality-sensitive hashing

The approximate nearest neighbor search problem is solved (efficiently) using hashing-based techniques that probabilistically group similar feature vectors together (see survey of Andoni *et al.* [12]). Approximate solutions based on Locality-sensitive Hashing (LSH) provide tunable accuracy guarantees and only require examining a small fraction of feature vectors in the database to find the approximate nearest neighbor(s).

LSH-based ANN search data structure

BUILD $(\mathcal{DB}, \mathcal{H}, L) \to (\mathcal{T}_1, \ldots, \mathcal{T}_L)$ takes as input a set of N vectors $\mathcal{DB} = \{\boldsymbol{v_1}, \ldots, \boldsymbol{v_N}\}$, LSH family $\mathcal{H}$ (Definition 1), and number of tables L. Outputs hash tables $\mathcal{T}_1, \ldots, \mathcal{T}_L$.

1: Sample L LSH functions $\mathsf{h}_1, \ldots, \mathsf{h}_L$ from $\mathcal{H}$.
2: Use h_i to build $\mathcal{T}_i$ by hashing each vector $\boldsymbol{v_1}, \ldots, \boldsymbol{v_N}$.
3: Output L hash tables $\mathcal{T}_1, \ldots, \mathcal{T}_L$.

QUERY $(\mathcal{T}_1, \ldots, \mathcal{T}_L, \mathsf{h}_1, \ldots, \mathsf{h}_L, \boldsymbol{q}) \to \mathsf{ID}$ takes as input a query vector $\boldsymbol{q}$, L hash tables, and LSH functions.

1: Compute bucket key $l \leftarrow \mathsf{h}_i(\boldsymbol{q})$ and retrieve the corresponding bucket $\mathcal{B}_l$ in $\mathcal{T}_i$ under key l (if non-empty).[a]
2: Set $\mathcal{C} := \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_L$.
3: Find j such that $\boldsymbol{v_j} \in \mathcal{C}$ and $\Delta(\boldsymbol{v_j}, cR) \leq cR$ via brute-force distance comparisons.
4: **if** no such j exists, output 0; **else** output j.

[a] Note that by the properties of LSH, the query will collide with probability proportional to the relative distance from other vectors.

Figure 2: ANN search data structure based on locality-sensitive hashing of Gionis et al. [55].

LSH families are defined over a distance metric (such as Euclidean distance) and have the property that feature vectors close to each other in space hash to the same value with good probability. Formally, for a fixed feature vector space $\mathcal{D}$, output space $\mathcal{R}$, and a distance metric Δ, an LSH family is defined as follows.

Definition 1 (Locality-sensitive Hash (LSH) Family)**.** A family of hash functions $\mathcal{H} := \{\mathsf{h} : \mathcal{D} \to \mathcal{R}\}$ is (R, cR, p_1, p_2)-sensitive for distance metric Δ if for any pair of vectors $\boldsymbol{v}, \boldsymbol{q} \in \mathcal{D}$,

$$\textbf{if } \Delta(\boldsymbol{v}, \boldsymbol{q}) \leq R \textbf{ then } \Pr[\mathsf{h}(\boldsymbol{v}) = \mathsf{h}(\boldsymbol{q})] \geq p_1,$$
$$\textbf{if } \Delta(\boldsymbol{v}, \boldsymbol{q}) \geq cR \textbf{ then } \Pr[\mathsf{h}(\boldsymbol{v}) = \mathsf{h}(\boldsymbol{q})] \leq p_2,$$

where $R < cR$ and $p_1 > p_2$.

Remark 1. Note that an LSH family is usually combined with a universal hash function to map to a fixed output size [44]. Without loss of generality, we assume that the output of the LSH function is mapped by a universal hash.

LSH for nearest neighbor search. In this work we adapt the data structure of Gionis et al. [55], which is the standard way of solving the ANN search problem using LSH (see survey of Andoni et al. [12]). The data structure consists of two algorithms: BUILD and QUERY, described in Figure 2.

Because the probability that a nearest neighbor collided with the query in a subset of hash tables can be made arbitrarily

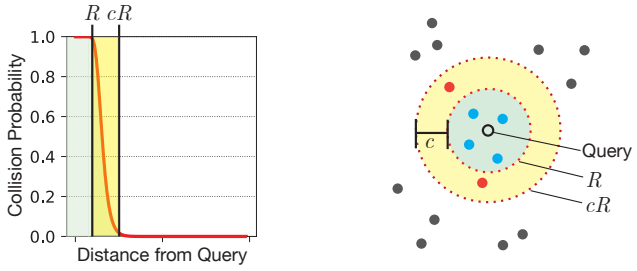

Figure 3: Visualization of the nearest neighbor search problem. **Left:** collision probability of an LSH-based data structure as a function of the distance between the query and points in the database. **Right:** points colliding with the query in the ANN data structure. Points within distance R have a high probability of colliding with the query whereas points within distance R and cR from the query have a lower probability of colliding. The approximation factor $c > 1$ determines the quality of the results; typically $c = 2$ in practical applications.

high (by tuning parameters), BUILD and QUERY ensure that the nearest neighbor is found with high probability (see Figure 3). In practice, one must query $L \approx \sqrt{N}$ tables to find a near neighbor in expectation [12].

3 Overview

We adapt the data structure of Section 2 into a privacy-preserving protocol between a client with query vector $\boldsymbol{q}$ and two servers with access to replicas of the database. We begin by describing the ideal functionality of private ANN search.

Notation. We denote by $\mathcal{DB}$ the database of vectors and their IDs. We let N be the total number of d-dimensional vectors in $\mathcal{DB}$. A vector is denoted in bold as $\boldsymbol{v}$ where the ith coordinate of $\boldsymbol{v}$ is denoted by $\boldsymbol{v}_i$. We denote the ith value of a set S by $S(i)$. A distance metric (e.g., Euclidean distance) is denoted Δ, where threshold distances R and cR are as defined in Section 2.1. We let $\mathbb{F}$ denote any prime-order finite field (e.g., integers mod a prime p). A secret-share of a value $v \in \mathbb{F}$ is denoted using bracket notation as $[v]$. Coordinate-wise secret-shares of a vector $\boldsymbol{v}$ is denoted $[\boldsymbol{v}]$. Variable assignment is denoted by $x \leftarrow y$, where $x \xleftarrow{R} S$ denotes a random sample from the set.

3.1 Ideal functionality

The ideal ANN search functionality is described in Functionality 1. The functionality takes as input the public parameters and query $\boldsymbol{q}$ to output the ID of the nearest neighbor relative to $\boldsymbol{q}$ to the client. The servers obtain no output. Without loss of generality, we assume the ID of the ANN is the *index* of the ANN for some canonical ordering of the feature vectors in the database. We let ID = 0 when no nearest neighbor exists.

Restricting the problem. Note that an LSH-based algorithm will only return an answer that is within distance cR of the query. We formalize this by assuming that the ideal functionality outputs the nearest neighbor that is also a *near*[2] (distance less than cR away from the query) neighbor. While it is possible to imagine contrived databases where the nearest neighbor is *not* also a near neighbor, most practical instances of the problem impose this additional restriction (returning no neighbor if the nearest neighbor is beyond a threshold distance from the query) because points beyond some threshold are effectively unrelated to the query. To this end, the ideal functionality is defined to reveal the ID of the nearest neighbor (if one exists) within a fixed distance $R = R_{\text{max}}$ from the query. Following Bayer *et al.* [25], we define two quantities D_{max} and D_{min} to be the maximum and minimum distance between any two points, respectively. Because the distance between any two vectors is at most D_{max}, it suffices to have $R_{\text{max}} < D_{\text{max}}$.

Functionality 1: Private ANN Search

Public Parameters:
- distance metric Δ.
- database size N.
- feature vector dimensionality d.
- maximum nearest neighbor radius R_{max}.

Inputs:
- Client: query $\boldsymbol{q} \in \mathbb{R}^d$.
- Server(s): database $\mathcal{DB} := \{\boldsymbol{v_1}, \ldots, \boldsymbol{v_N} \mid \boldsymbol{v_i} \in \mathbb{R}^d\}$.

Procedure:
1: $\boldsymbol{v_j} \leftarrow$ nearest neighbor to $\boldsymbol{q}$ in $\mathcal{DB}$ via brute-force search.
2: **if** $\Delta(\boldsymbol{v_j}, \boldsymbol{q}) > R_{\text{max}}$ **then**
output 0 to the client and $\perp$ to the servers.
3: **else** output j to the client and $\perp$ to the servers.

3.2 Threat model and security guarantees

Our protocol is instantiated with two non-colluding servers and an arbitrary number of clients. Clients query the servers to obtain an ID of the ANN. We do not require any communication between servers when answering queries.

Threat model.

- No client is trusted by either server. Clients may deviate from protocol, collude with other clients, or otherwise behave maliciously to learn more about the database.
- No server is trusted by any client. One or both servers may deviate from protocol in an attempt to obtain information on a client's query or the resulting nearest neighbor.

Assumptions. Our core assumption, required for client privacy, is that the two servers do not collude with each other. Additionally, we require black-box public-key infrastructure (e.g., TLS [100]) to encrypt communication over the internet.

[2]This name comes from the *near* neighbor search problem [63][6].

Guarantees. Under the above threat model and assumptions, the protocol provides the following guarantees.

Correctness: If the client and servers both follow protocol, then the client obtains the ID of the ANN with respect to its query. The result is guaranteed to have the same approximation accuracy of standard, non-private data structures for ANN search, and has tunable accuracy guarantees.

Client privacy: If the servers do not collude, then neither server learns any information on the client's query, even if one or both servers arbitrarily deviate from protocol.

Bounded leakage: Each query answer is guaranteed to leak a small (and tightly bounded) amount of information over the ideal functionality, even if the query is maliciously generated by the client. We provide a precise definition and in-depth analysis of this leakage in Section 7.

4 Main ideas

To introduce privacy, as required for the client *and* the database, we make several changes to the data structure of Figure 2 (BUILD and QUERY). A simple strawman protocol with *client* privacy (but no database privacy) can be realized by applying well-known techniques in Private-information Retrieval (PIR) to privately obtain the answer to QUERY (see [38, 39, 52]). PIR allows a client to privately retrieve a specified object from a remote database without revealing which object was retrieved, which naturally generalizes to retrieving buckets from a hash table [39]. While PIR solves the client privacy problem, it provides no database privacy. The client learns $\sqrt{N}$ (see Section 2.1) feature vectors from the database *per query*.

The challenge with database privacy. The primary challenge in reducing database leakage comes from preventing the client from learning extra vectors in the candidate set. This is non-trivial to do given that the standard approaches to removing false-positive candidates (points farther than cR from the query) require some form of direct distance comparisons. In the private setting, these become *oblivious* comparisons (a comparison between secret-shared values), which in turn require *heavy* cryptographic techniques (e.g., garbled circuits [116]). The state-of-the-art approach for privacy-preserving ANN search (SANNS [36]) prunes candidates by using an expensive two-party computation, which requires several gigabytes of communication between the client and database server.

The insight that we exploit to overcome this challenge is that LSH can be used directly to accomplish the same goal of pruning false-positives. We then apply a trick inspired by radix sorting [72] to extract the *nearest* neighbor, fully removing the need for direct comparisons between vectors. Our new data structure is slightly less efficient when viewed from an *algorithmic* perspective (i.e., when compared to the data structure of Figure 2). However, this is not a problem for us given that oblivious comparisons are the primary bottleneck in a privacy-preserving setting. We elaborate on this observation in the next section.

4.1 Reframing the problem

LSH-based ANN search is typically optimized to minimize the number of hash tables (L) and the size of the candidate set for each query. Removing false-positives via brute-force comparisons is relatively "cheap" from a computational standpoint while hash table lookups are relatively expensive. Therefore, LSH-based ANN search is typically tuned to retrieve as many (reasonable) candidates as possible from each table. The extra candidates are then pruned via brute-force comparisons.

The privacy-preserving setting requires different priorities. First, note that it is not possible to perform only one lookup per hash table. To preserve privacy, *all* hash table buckets must be "touched" by the database server(s) to avoid revealing information on the client's query. This is the lower bound on private information retrieval [56]. Otherwise, the servers learn that the client's hash does *not* correspond to any untouched bucket. This is exactly why existing solutions for privacy-preserving ANN search require $O(N)$ communication between the client and server, or alternatively, the use of fully-homomorphic encryption (which requires linear server work). As such, we cannot hope to have sublinear (in N) work for the servers when answering client queries.

With this observation in mind, we note that a sublinear time protocol is not possible, and realize that the optimal LSH parameters in the non-private setting might not in fact be optimal for privacy-preserving data structures. We therefore redesign and re-tune the ANN search data structure of Section 2 for the purpose of avoiding comparisons between vectors.

4.2 Eliminating oblivious comparisons

Goal: No false-positives. One idea to remove comparisons is to prevent values that will be pruned from being added to the candidate set in the first place. More precisely, by tuning the parameters of the ANN search data structure, and the LSH functions it uses, we can bound the probability of false-positives in the candidate set to any $0 < \delta < 1$.

4.2.1 Bounding false-positives

A standard result in LSH theory is that a (R, cR, p_1, p_2)-sensitive hash family can be "amplified" to result in a (R, cR, p_1', p_2')-sensitive hash family, where $p_1' = p_1^k$ and $p_2' = p_2^k$. This is done by simply concatenating the outputs of k independent (R, cR, p_1, p_2)-sensitive hash functions. Amplification is generally used to reduce the size of the candidate set [12], however, we take this approach to its extreme. For a larger value of k, fewer collisions will occur. On the other

hand, when considering collisions that do occur, they are more likely to be true positives. As a result, with a sufficiently large k, we can significantly reduce false positives and compensate for the smaller p_1' by increasing the number of tables L. We capture this idea in Proposition 1.

Proposition 1. Fix failure probability $\delta > 0$, and any (p_1, p_2, R, cR)-sensitive hash family where $\rho = \frac{\log(p_1)}{\log(p_2)} < \frac{1}{2}$. There exists a data structure solving the cR-approximate near neighbor problem in $O(N^{\rho'})$ time and $O(N^{1+\rho'})$ space, where $\rho < \rho' < 1$. This data structure returns the true cR-approximate nearest neighbor with probability $1-\delta$, without requiring brute-force distance comparisons between vectors.

Proof. The LSH data structure of Section 2.1 (Figure 2) achieves asymptotic space and query time $O(N^{1+\rho})$ and $O(N^{\rho})$, respectively [12]. Consider the data structure described in Figure 2 where QUERY is modified to a *random* element $\in \mathcal{C}$ (if $\mathcal{C} \neq \emptyset$) instead of an element that is guaranteed to be within distance cR. We need to bound the success probability of this random element being within cR of the query to $1-\delta$. By Definition 1 (LSH), we have that $p_1 > p_2$. Therefore, there exists a "gap" between the probability of a true positive and a false positive collision, which can be amplified exponentially in k. First, observe that the probability of selecting a false-positive at random from $\mathcal{C}$, when the database size is N, is bounded by

$$\frac{N \cdot p_2^k}{p_1^k + N \cdot p_2^k} < N \cdot \frac{p_2^k}{p_1^k}, \tag{1}$$

which we further need to bound by δ. Since we have that $p_1 > p_2$ (Definition 1), it follows that $p_2/p_1 < 1$. Therefore, when

$$k \geq \left\lceil \frac{\log(N) + \log(1/\delta)}{\log(1/p_2) + \log(p_1)} \right\rceil$$

we get that (1) is bounded by δ, which leads to success probability $1-\delta$ of selecting a true-positive at random (when it is contained in the candidate set). We contrast this to "vanilla" LSH-based data structures (e.g., Figure 2) where $k \geq \left\lceil \frac{\log(N)}{\log(1/p_2)} \right\rceil$ [12]. We note that the values are asymptotically similar. The new data structure results in the elimination of brute-force comparison while still preserving accuracy guarantees. Finally, to prove space and query time, we note that the expected number of tables is

$$L = \lceil p_1^{-k} \rceil = N^{\frac{\rho}{1-\rho}} \cdot \delta^{\frac{1-\rho}{\rho}}.$$

Therefore, we get $\rho' < 1$ when $\rho < \frac{1}{2}$. Following the LSH-based ANN search data structure of Section 2, the asymptotic query time is $O(L)$ and the space is $O(N \cdot L)$. ∎

When is $\rho < \frac{1}{2}$? The restriction $\rho < 1/2$ (required for sublinearity in Proposition 1) is met for many common distance metrics when the approximation factor $c \geq 2$. We cite various results for when these conditions are met. We note that these values are for *worst-case* theoretical guarantees; in practice, ρ is much smaller (see Section 8). For Euclidean distance, Andoni and Indyk [8] show $\rho < \frac{1}{c^2} = \frac{1}{4}$ with $c = 2$. Andoni et al. [10] also show a similar result for angular distance. For any p-stable distribution with $1 \leq p \leq 2$, Datar et al. [44] show $\rho < \frac{1}{c} = \frac{1}{2}$ for $c = 2$. For the ℓ_1-norm specifically, Motwani et al. [86] show $\rho < \frac{1}{2c} = \frac{1}{4}$ with $c = 2$. Hamming distance can be embedded into Euclidean space (see Aumüller et al. [17]). Alternatively, [117, Corollary 3.10] shows $\rho < \frac{1}{c} = \frac{1}{2}$ for hamming distance directly.

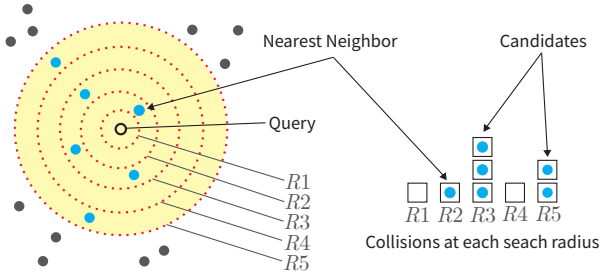

Figure 4: Illustration of using multiple radii (R) to search for the nearest neighbor. **Left:** Each of the dotted regions represents a different hash function radius. **Right:** The candidate result with the smallest R_i is the nearest neighbor, in this case the bucket corresponding to R_2.

Bounded false-positives (in the worst case). The consequence of Proposition 1 is that while we can bound false-positives to any δ, this comes with the cost of increasing the number of tables L, since $L = \lceil p_1^{-k} \rceil$. Because k is a function of the LSH sensitivity, we need to ensure that the difference between p_1 and p_2 is sufficiently large to result in reasonable values of k and L. We describe such LSH families in Appendix A. In our evaluation (Section 8), we show that on real data, we can have false positive probability less than 0.05 with $k = 2$. With Proposition 1, we can guarantee that all collisions are within cR from the query, with high probability.

Finding the nearest neighbor. We are now left with the problem of finding the *nearest* neighbor within the set of all cR-neighbors. Our idea for doing so is based on a bucketing technique of Ahle *et al.* [63], which resembles radix sorting [72]. A radix sort does not perform direct comparisons, which aligns with our goals. We recursively apply the data structure of Proposition 1 on a series of increasing neighbor radii, retrieving a set of candidates from each radius [63]. The client then selects a result from the first non-empty candidate set (see Figure 4).

4.2.2 The privacy-friendly ANN data structure

Our new ANN data structure (presented in Figure 5) merges the above ideas to eliminate the brute-force step present in the

Comparison-free ANN search data structure

BUILD $(\mathcal{DB}, \mathcal{H}_1, \ldots, \mathcal{H}_L) \to (\mathcal{T}_1, \ldots, \mathcal{T}_L, \mathsf{h}_1, \ldots, \mathsf{h}_L)$ takes as input a set of N vectors $\boldsymbol{v_1}, \ldots, \boldsymbol{v_N}$ and L hash families $\mathcal{H}_i$ each corresponding to an $R_L \le R_i \le R_L$ (Section 4.2).

1: Sample random h_i from $\mathcal{H}_i$, for $i \in \{1, \ldots, L\}$.
2: Use h_i to build $\mathcal{T}_i$ by hashing each vector $\boldsymbol{v_1}, \ldots, \boldsymbol{v_N}$.
3: Output the similarity search data structure consisting of L hash tables $\mathcal{T}_1, \ldots, \mathcal{T}_L$ and LSH functions $\mathsf{h}_1, \ldots, \mathsf{h}_L$.

QUERY $(\mathcal{T}_1, \ldots, \mathcal{T}_L, \mathsf{h}_1, \ldots, \mathsf{h}_L, \boldsymbol{q}) \to \mathsf{ID}$; as in Figure 2.

1: Compute $l_i \leftarrow \mathsf{h}_i(\boldsymbol{q})$ and retrieve bucket $\mathcal{B}_{l_i}$ from $\mathcal{T}_i$.
2: $j \leftarrow \min i$ with nonempty $\mathcal{B}_i$, or 0 if no such i exists.
3: **if** $j \neq 0$ **then** output any γ s.t. $\boldsymbol{v_\gamma} \in \mathcal{B}_j$; **else** output 0.

Figure 5: ANN search data structure with no direct comparisons.

vanilla LSH-based data structure of Figure 2. The client can simply select any element from the first non-empty candidate set. This data structure can be further adapted to suppress database leakage, as we explain in the next section.

4.3 Database privacy and suppressing leakage

In this section we explain how the ideas of Section 4.2 apply to suppress database leakage. Our approach is a combination of three changes applied to QUERY in Figure 5. We recall that a simple strawman protocol achieving client privacy can be constructed by having the client privately retrieve colliding buckets through PIR [39].

Step 1: Capping buckets. Because there is only one element to retrieve in expectation from the data structure of Figure 5, we can limit each hash bucket to only contain one vector. This ensures the size of the candidate set never exceeds $\sqrt{N}$.

Step 2: Hiding the vectors. Because the client selects the first non-zero ID from the candidate set using the data structure described in Figure 5, it does not need access to the vectors. As such, we can modify each hash table (BUILD; Figure 5) to only store the IDs of each vector. The client can still query the hash tables using PIR but now only obtains a candidate set of IDs (absent the vectors). If each vector in the database is d dimensional, then this simple change reduces leakage from $O(\sqrt{N} \cdot \log N \cdot d)$ bits to $O(\sqrt{N} \cdot \log N + d)$ bits (each ID is at least $\log N$ bits and $O(d)$ bits are leaked *implicitly* by the inference that the neighbor features are similar to the query).[3]

Step 3: Hiding the candidate set. Compared to the ideal leakage of $O(\log N + d)$ bits, the leakage of $O(\sqrt{N} \cdot \log N + d)$ bits is far from optimal. We remove this extra leakage by designing a special "oblivious masking" transformation which hides all-but-one non-zero candidate ID from the client. From the masked candidate set $\tilde{\mathcal{C}}$, the client is only able to learn at most one ID that collided with its query. This further reduces leakage from $O(\sqrt{N} \cdot \log N + d)$ bits down to $O(\log N + d)$ bits (since only one ID is revealed), which matches the asymptotic leakage of the ideal functionality. We provide details on the oblivious masking transformation in Section 5.2, and a more formal leakage analysis in Section 7.

With these three leakage-suppressing steps, our protocol achieves close to optimal concrete database leakage per query. Importantly, the leakage guarantees hold in the face of malicious clients that may deviate from protocol in an attempt to learn more than an honest client.

[3] A more precise leakage analysis is found in Section 7.

5 Protocol

We now describe the details of the high-level ideas covered in Section 4. We first formalize the necessary building-blocks in Section 5.1 and Section 5.2 (distributed point functions and our oblivious masking technique) and then present the full ANN search protocol in Section 5.3.

5.1 Building blocks

Existing tool: Distributed Point Functions. A *point function* P_i is a function that evaluates to 1 on a single input i in its domain and evaluates to 0 on all other inputs $j \neq i$. A *distributed* point function (Definition 2) is a point function that is encoded into a pair of keys which are used to obtain a secret-shared evaluation of P_i on a given input j.

Definition 2 (Distributed Point Function (DPF) [29, 30, 54])**.** Fix any positive integer D. Let $\mathbb{F}$ be any finite field (e.g., integers mod prime p), and let λ be a security parameter. A DPF consists of two (possibly randomized) algorithms:

- $\mathsf{Gen}(1^\lambda, i \in \{1, \ldots, D\}) \to (k_A, k_B)$ takes as input an index i and outputs two evaluation keys k_A and k_B,
- $\mathsf{Eval}(k, j) \to v_j \in \mathbb{F}$ takes as input an evaluation key k and index $j \in \{1, \ldots, D\}$ and outputs a field element v_j.

These algorithms must satisfy correctness and privacy:

Correctness. A DPF is *correct* if for all pair of keys generated according to Gen,

$$\Pr\left[\ \mathsf{Eval}(k_A, j) + \mathsf{Eval}(k_B, j) = \begin{cases} 1, & \text{if } j = i \\ 0, & \text{otherwise.} \end{cases}\ \right] = 1,$$

Privacy. A DPF is *private* if each individual evaluation key output by Gen is pseudorandom (i.e., reveals nothing about i to a computationally bounded adversary). Formally, this means that there exists an efficient simulator Sim that can generate an indistinguishable view for each generated DPF key, without knowledge of the input i [29, 54].

Application: symmetric PIR. DPFs form the basis for efficient two-server PIR with optimal communication. To see this, consider an ordered list S of N elements replicated on two servers. To retrieve the index i in S the client generates DPF keys and sends a share to each server. The servers evaluate the DPF on each index $j = 1, \ldots, N$ to obtain a secret-share of the point function evaluated on input j. Each server then multiplies the resulting secret-share by the corresponding entry in S at index j, and outputs the sum of all N component-wise products. That is, each server computes:

$$\sum_{j=1}^{N} \Big(S(j) \cdot \mathsf{Eval}(k, j) \Big) = S(i) \cdot [1] + \sum_{j \neq i}^{N} \Big(S(j) \cdot [0] \Big).$$

Because $P_i(j) = 1$ only when $i = j$, the resulting sum (computed in the field $\mathbb{F}$) is a secret-share of $S(i)$. Summing the shares received from each server, the client recovers the desired entry in S. Observe that DPFs achieve *symmetric* PIR [53] (analogous to oblivious transfer [96]), which guarantees that the answer consists of at most one element in S.

Extension: PIR-by-keywords. The key size of modern two-server DPF constructions is $O(\log D)$ bits [29, 30]. This allows for PIR with $O(\log D + m)$ communication (where m is the size of elements in S). More generally, DPFs enable keyword-based PIR [39]. That is, for a key-value table consisting of N keys drawn from the domain $\{1, \ldots, D\}$, it is possible to use a DPF in a two-server setting to privately retrieve a value stored under key $i \in \{1, \ldots, D\}$, with only $O(\log D + m)$ communication and $O(N \cdot m)$ work on each server [29, 45]. The latter follows from the ability to "lazily evaluate" DPFs on only the N keys present in the key-value table, rather than the full domain D.

5.2 New tool: oblivious masking

The core of our leakage suppression technique (described at a high level in Section 4.3) hinges on the ability to reveal only the first nonzero value in a secret-shared vector. Because the vector is secret-shared (in some prime order field $\mathbb{F}$), this transformation must only involve affine operations: addition and scalar multiplication of shares [24, 103]. The idea is to recursively compute a randomized sum, moving from left to right. For a secret-shared input vector $[\boldsymbol{v}] \in \mathbb{F}^t$, let $z \in \{1, \ldots, t\}$ be the index of the first nonzero element in $\boldsymbol{v}$. The randomized sums map each element $\boldsymbol{v}_i$ to 0 for $i < z$ and a uniformly random element in $\mathbb{F}$ for $i > z$. Crucially, this process does not affect the first non-zero element, $\boldsymbol{v}_z$. We will use this property to mask all-but-one result from a sequence of PIR query answers.

Claim 1. Let $\boldsymbol{v} \in \mathbb{F}^t$ be any vector and let $z \in \{1, \ldots, t\}$ be the first non-zero element of $\boldsymbol{v}$. Let $[\boldsymbol{v}]$ be an additive secret-sharing of $\boldsymbol{v}$. Algorithm 1 outputs a secret-shared vector $[\boldsymbol{y}]$ such that $\boldsymbol{y}_i = \boldsymbol{v}_i$ for $i \leq z$ and a uniformly random element of $\mathbb{F}$ for $i > z$.

Algorithm 1: ObliviousMasking

Input:

- Secret-shared vector $[\boldsymbol{v}] \in \mathbb{F}^t$.
- Randomness source rand.

Output:

- Secret-shared vector $[\boldsymbol{y}] \in \mathbb{F}^t$.

Procedure:

1: **for** $i \in \{1, \ldots, t\}$**:**

1.1: Sample $r_i \xleftarrow{R} \mathbb{F}$ according to rand.

1.2: Set $[\boldsymbol{y}_i] \leftarrow [\boldsymbol{v}_i] + \boldsymbol{r}_i \cdot \Big(\sum_{j=0}^{i-1} [\boldsymbol{v}_j] \Big)$.

2: Output $[\boldsymbol{y}] \in \mathbb{F}^t$.

Proof. The proof follows by examining the three possible cases for each value in $\boldsymbol{y}$ as a function of $\boldsymbol{v}$.

1. for $i < z$, $\boldsymbol{v}_i = 0$, so $\boldsymbol{y}_i = 0 + r_i \sum_{j=0}^{i-1} 0 = 0 \in \mathbb{F}$,
2. for $i = z$, $\boldsymbol{y}_z = \boldsymbol{v}_z + r_z \sum_{j=0}^{z-1} 0 = \boldsymbol{v}_z \in \mathbb{F}$,
3. for $i > z$, $\boldsymbol{y}_i = \boldsymbol{v}_i + r_i \sum_{j=0}^{z-1} 0 + r_i \boldsymbol{v}_z + r_i \sum_{j=z+1}^{i-1} \boldsymbol{v}_j \in \mathbb{F}$.

Case (1) ensures that all zeroes remain zeroes. Case (2) ensures that the first non-zero element is mapped to itself. Case (3) ensures that all subsequent elements are uniformly random in $\mathbb{F}$. To see why (3) holds, observe that $\boldsymbol{v}_z \neq 0$, so $r_i \cdot \boldsymbol{v}_z$ is a uniformly random element of $\mathbb{F}$ given r_i is uniformly random. It then follows that the sum is uniformly random in $\mathbb{F}$. Finally, correctness of the computation over secret-shares follows from all operations performed above being linear (additions and scalar multiplications) over the input secret-share of $\boldsymbol{v}$ [24]. ■

5.3 Putting things together

The full protocol is presented in Protocol 1 and uses the DPF and oblivious masking building-blocks described in Section 5.1 and Section 5.2. We briefly describe each step of the protocol.

Setup. The public parameters consist of the number of hash tables (L) and a list of L randomly sampled hash functions, in accordance with the data structure of Figure 5 and Proposition 1. The servers construct L hash tables using the hash functions from the public parameters. Only the IDs (i.e., indices) of the input vectors are stored in the hash table; the vectors are discarded (see Section 4.3).

Step 1. The client hashes its query vector $\boldsymbol{q}$ using the LSH functions in the public parameters. Each resulting hash is used as the keyword in DPF.Gen to generate a DPF key (see PIR-by-keywords extension described in Section 5.1). The generated DPF keys are distributed by the client to each server.

Step 2. Each server uses the L DPF keys it receives from the client to retrieve a secret-share of a bucket in each of the L hash tables using PIR-by-keywords. The result is a list $\mathcal{C}$ of L

secret-shared buckets containing either a candidate ID (or zero if the bucket was empty). Each server applies the oblivious masking transformation (Algorithm 1) to $\mathcal{C}$ and obtains the masked secret-shared list $\tilde{\mathcal{C}}$ as output. Each server sends its share of $\tilde{\mathcal{C}}$ to the client in response.

Step 3. From the received secret-shares $(\tilde{\mathcal{C}}_A, \tilde{\mathcal{C}}_B)$, the client recovers the list $\tilde{\mathcal{C}}$ (in cleartext) by computing $\tilde{\mathcal{C}} = \tilde{\mathcal{C}}_A + \tilde{\mathcal{C}}_B \in \mathbb{F}^L$.[4] The transformation applied by the servers in Step 2 ensures that the client learns at most *one* non-zero candidate from the original $\mathcal{C}$. Specifically, by Claim 1, only the first non-zero value of $\mathcal{C}$ will be non-random in $\tilde{\mathcal{C}}$. The first non-zero value, if present, is output to the client.

Concrete efficiency. While Protocol 1 satisfies our asymptotic efficiency requirements, it requires a few more optimizations to achieve concrete efficiency. Specifically, we show how to optimize query processing time on the server using hash table multi-probing [82, 93] and a new probabilistic batch-processing technique [13, 14, 66] for PIR. We present these optimizations to the protocol in Section 6.

5.4 Asymptotic efficiency analysis

We analyze the communication and computation costs individually (summarized in Table 1). To derive our asymptotic guarantees, we follow the analysis of Andoni and Indyk [8] for Euclidean distance (which has $\rho < \frac{1}{c^2}$; see definition of ρ in Proposition 1) and where we assume $c = 2$ (see Definition 1).

Communication	Server Work	Client Work	Rounds
$\sqrt{N}$	N	$\sqrt{N}$	1

Table 1: Asymptotic efficiency of Protocol 1; constant and log factors suppressed for readability. We assume $L \approx \sqrt{N}$ (see Section 2.1).

Communication. The protocol communication consists of the L DPF keys, each of which retrieves one candidate from a table. The DPF key size is $O(\log N)$ [29, 30]. Each masked ID returned by the protocol is of size $\log N$. Therefore, when $L \approx \sqrt{N}$, the total communication is $O(\sqrt{N} \log N)$.

Computation. For each server, computing the answer to the PIR query requires evaluating the DPF on $O(N)$ hash keys per hash table. Therefore, the total server work is naïvely $O(L \cdot N)$. In Section 6.2, we show a new technique that we use to amortize this processing cost to $O(N)$ by retrieving multiple candidates from each hash table simultaneously. Oblivious masking, which is performed over the candidate set, requires $O(\sqrt{N})$ additional work per server. For the client, the computational cost is proportional to the query response size, since it has to receive and recover the candidate set and find the first non-zero element to return as the answer.

[4] Additive secret-share recovery follows from the correctness property of DPFs; see Definition 2.

Protocol 1: Private Similarity Search

Public input: LSH family $\mathcal{H}$, number of hash tables L, and LSH functions $\mathsf{h}_1, \ldots, \mathsf{h}_L$ sampled i.i.d. from $\mathcal{H}_i$.
Server input: database of vectors $(\boldsymbol{v_1}, \ldots, \boldsymbol{v_N})$.
Client input: query vector $\boldsymbol{q}$.

Setup (one-time server-side pre-processing)
On each server

1: **for** $i \in \{1, \ldots, L\}$**:**

1.1: Construct hash table $\mathcal{T}_i$ by storing j in bucket with key $l_j := \mathsf{h}_i(\boldsymbol{v_j})$ for all $j \in N$.

1.2: Truncate each bucket in $\mathcal{T}_i$ have at most one value; discarding values at random.

2: Agree on common randomness source rand.

Protocol

Step 1:
On the client

1: **for** $i \in \{1, \ldots, L\}$**:**

1.1: $l_i \leftarrow \mathsf{h}_i(\boldsymbol{q})$. `// get hash key for table` $\mathcal{T}_i$.

1.2: $(k_A^{(i)}, k_B^{(i)}) \leftarrow \mathsf{Gen}(l_i)$. `// DPF keys for` $\mathcal{T}_i$ `query.`

2: $\boldsymbol{k}_A \leftarrow (k_A^{(1)}, \ldots, k_A^{(L)})$ and $\boldsymbol{k}_B \leftarrow (k_B^{(1)}, \ldots, k_B^{(L)})$.

3: Send $\boldsymbol{k}_A$ and $\boldsymbol{k}_B$ to servers A and B, respectively.

Step 2:
On each server

1: Parse $\boldsymbol{k} = (k^{(1)}, \ldots, k^{(L)})$. `// set of DPF keys.`

2: **for** $i \in \{1, \ldots, L\}$**:**

2.1: $n_i \leftarrow$ number of bucket keys $(l_1, \ldots, l_{n_i})$ in hash table $\mathcal{T}_i$.

2.2: $[a_i] \leftarrow \sum_{j=1}^{n_i} \mathcal{T}_i(l_j) \cdot \mathsf{Eval}(k^{(j)}, l_j)$. `// Keyword PIR.`

3: $\mathcal{C} \leftarrow ([a_1], \ldots, [a_L])$.

4: $\tilde{\mathcal{C}} \leftarrow$ ObliviousMasking$(\mathcal{C}, \mathsf{rand})$. `// Algorithm 1.`

Step 3:
On the client

1: Receive $\tilde{\mathcal{C}}_A$ and $\tilde{\mathcal{C}}_B$ from servers A and B, respectively.

2: **for** $i \in \{1, \ldots, L\}$**:**

2.1: $j_i \leftarrow \tilde{\mathcal{C}}_A(i) + \tilde{\mathcal{C}}_B(i)$. `// recover secret-shared ID.`

3: Output $j_i \neq 0$ for smallest i, or 0 if all j_i are zero.

6 Optimizations and extensions

We now turn our attention to making the protocol concretely efficient. The primary bottleneck in Protocol 1 comes from the processing time needed to retrieve L buckets using the provided DPF keys. While this processing time does not involve the client, it can lead to high latency in answering queries, as each query requires $O(N)$ server-side work per hash table.

The seminal work of Ishai et al. [67] provides a way to amortize the processing cost of performing multiple PIR queries

using *batch codes*. Batch-PIR [67] computes the answer to ℓ simultaneous PIR queries at an amortized processing cost. While the idea of batching queries is appealing, it does not directly apply to Protocol 1. The reason is that batch-PIR only works when retrieving multiple elements from the *same* table; not when querying different tables. To align with the requirements of batch-PIR, we use a technique known as LSH multi-probing [82, 93]. Multi-probing retrieves multiple keys in each hash table and reduces the total number of hash tables in the ANN data structure. Then, we can use batching techniques to retrieve all multi-probes from each table efficiently. However, we find that the traditional way of performing batch-PIR through coding schemes results in too much overhead in probabilistic settings. To remedy this, we develop a new tool we call *partial batch retrieval*. We show that combining these two techniques results in several orders of magnitude faster processing time.

6.1 LSH multi-probing

The idea behind LSH multi-probing is the following: if the bucket to which the query hashes to in a table is empty, then it is likely that "adjacent" buckets in the table contain the collision [82, 93]. To exploit this observation, each hash table is probed on ℓ keys which "surround" the query. The surrounding keys are computed by hashing a perturbed version of the query vector, leading to a higher likelihood of finding a candidate, if it exists. In turn, this reduces the number of hash tables (L) required (the original motivation of multi-probing [82]). The appealing property to our use case, however, is that multi-probing allows us to apply batch-PIR techniques and heavily amortize processing time when retrieving candidates. The only change to Protocol 1 is that now the client retrieves up to ℓ candidates from each hash table, instead of just one candidate. This reduces the number of hash tables (L) required by a factor of ℓ. Meanwhile, answering a batch of ℓ queries can be amortized to approximately the cost of answering one query, which we show next.

6.2 Partial batch retrieval

Batch codes [66] have the strict requirement that *any* combination of ℓ elements should be retrievable with perfect probability. With Partial Batch Retrieval (PBR) (inspired by *probabilistic* batch codes [13, 14]), we relax this requirement: we only care about retrieving some *fraction* of the ℓ elements, with high probability.

The main idea. A simple PBR can be realized by dividing the elements into $m \geq \ell$ partitions at random. If all ℓ elements a client would like to retrieve fall into a unique partition, then it suffices for the client to issue m PIR queries, with each query retrieving an element from a partition. The total server processing time to compute PIR answers remains the same. To see this, observe that for a database of N elements, each partition contains N/m elements. The total work to answer m PIR queries, each computed over a set of N/m elements, is then amortized to $O(N)$.

Batch-PIR scheme	Replication Factor	Partitions	Fraction Retrieved
Naive	1	1	$\frac{1}{\ell}$
SBC [66]	$(\frac{3}{2})^{\ell}$	$3^{\log \ell}$	1
PBC [14]	3	1.5ℓ	$> 1 - 2^{-20}$
PBR	1	m	$\frac{m}{\ell} \cdot (1 - e^{-\ell/m})$

Table 2: Replication and partitioning costs of existing batching schemes: Naive (perform ℓ PIR queries), Subcube batch codes (SBC) [66], Probabilistic batch codes (PBC) [14]. Replication increases server-side processing by the same factor but influences communication by a sublinear factor. Communication increases linearly with the number of partitions for all schemes.

However, the success of this PBR scheme hinges on elements falling into unique partitions. How many elements can we expect to retrieve in practice?

Abstractly, the fraction of retrievable elements can be modeled by the classic "balls and bins" problem [85], where ℓ balls are tossed into m bins uniformly at random. If all ℓ balls fall into unique bins, then the number of full bins is ℓ. Fewer than ℓ full bins corresponds to a collision in a bin (i.e., partition) and only one element in each partition can be retrieved. Let X_i be the indicator random variable where $X_i = 1$ if a bin is full. Then,

$$\Pr[X_i = 1] = 1 - \left(1 - \frac{1}{m}\right)^{\ell} > 1 - e^{-\ell/m},$$

which implies that $\frac{m}{\ell} \cdot (1 - e^{-\ell/m})$ of the ℓ elements are simultaneously "retrievable" from the $m \geq \ell$ partitions. In the case that $m = \ell$, we can expect to retrieve approximately 63% of the elements. With $m > \ell$, we can increase the probability of retrieving all elements at the cost of also increasing communication by a factor of $\frac{m}{\ell}$.

Comparison to batch codes. It is natural to ask how PBRs compare to batch codes. The main difference is with respect to the guarantees provided by both tools. Batch codes (and their probabilistic variants [13, 14]) aim to guarantee retrieval of *all* ℓ elements. Doing so comes at an efficiency cost: batch codes *replicate* data which increases both processing time and bandwidth. With PBRs, we avoid the need for replicating data, keeping the total *processing* cost fixed while only modestly increasing bandwidth (see Table 2 and Figure 6 for a comparison to batch codes). However, this comes at a different cost: PBRs only guarantee *partial* retrieval of the ℓ elements. We find that for certain applications, such as Protocol 1, partial retrieval in favor of decreased processing time is a desirable trade-off, given the already probabilistic guarantees of LSH.

Application to Protocol 1. We apply the above PBR scheme to retrieve a fraction of multi-probe hash keys from each hash

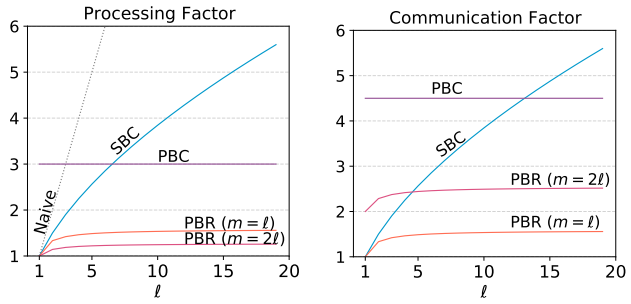

Figure 6: Concrete overhead factor, in terms of communication and processing time, for retrieving ℓ elements using our partial batch retrieval (PBR; Section 6.2), subcube batch codes (SBCs [66]), and probabilistic batch codes (PBCs [14]).

table in Protocol 1. Our main observation is that failing to retrieve specific probes is not a total failure. It is equivalent to not choosing that particular multi-probe, which could have already happened due to the probabilistic nature of LSH. Because each multi-probe hash key is uniformly distributed from universal hashing (see Remark 1), each hash key is equally likely to be selected, making it possible to directly apply our PBR scheme. The use of a PBR makes Protocol 1 both concretely and asymptotically efficient. The number of hash tables L decreases by a factor proportional to ℓ, amortizing the processing cost for both servers. Setting $m = \ell = L/a$, for some constant a, results in total amortized processing time that is $O(N)$ (as referenced in Section 5.4), which matches the PIR processing lower bound [38].

7 Security and leakage analysis

In this section, we analyze the security of Protocol 1 with respect to client privacy and server leakage. While client privacy is conceptually simple and follows directly from the privacy property of DPFs (Definition 2), the leakage analysis of the database is more involved.

7.1 Client privacy

Claim 2 (Query Privacy)**.** For all Probabilistic Polynomial Time (PPT) adversaries $\mathcal{A}$ corrupting server $\circ$ for $\circ \in \{\mathrm{A}, \mathrm{B}\}$, Protocol 1 guarantees that $\mathcal{A}$ learns no information on the client's query, even when deviating from protocol. This holds even when the query-batching optimization (Section 6) is applied to Protocol 1.

Proof. The proof follows from a simulation-based indistinguishability argument. We say that the protocol is *query private* if there exists a PPT simulator Sim such that

$$\mathsf{Sim}(\mathcal{DB}) \approx_c \mathsf{view}(\mathcal{DB}, \boldsymbol{k}_\circ),$$

where view is the view of $\mathcal{A}$ from an execution of Protocol 1 with the client. $\mathsf{Sim}(\mathcal{DB})$ is trivially constructed by invoking the DPF simulator for each DPF key present in the vector $\boldsymbol{k}_\circ$ provided to $\mathcal{A}$ [32]. The client does not send any other information to the servers apart from the DPF keys which are used to query the hash tables. Therefore, query privacy depends only on the privacy property of DPFs and follows immediately.

With multi-probing and batching (optimizations described in Section 6), the client retrieves ℓ keys from each table, where the key space is partitioned into m uniformly random subsets (see Section 6.2). This results in m DPF keys sent to each server per hash table. Because the client sends a DPF key for each partition (even if no bucket is retrieved from that partition), no information is revealed to $\mathcal{A}$. Therefore, query privacy follows by the same argument. See also Ishai et al. [66]. ∎

7.2 Database privacy and leakage

We now turn to formalizing the database leakage incurred when answering a client query. We must be careful to define database leakage in a meaningful way. As observed by Zuber and Sirdey [119]: *"[H]aving access to the result of the classification leads to some (useful) leakage on the raw database."* As such, we must start from a *baseline leakage* given that the client obtains information as output: the ANN ID relative to the query $\boldsymbol{q}$. In our setting, the client learns strictly more than just the ID of the ANN, which we must quantify as additional leakage. We emphasize that this leakage is with respect to the server's database, **not** the client's query.

Theorem 1 (Quantifying Ideal Leakage)**.** Fix quantities D_{max}, D_{min}, and Δ as defined in Section 3.1. The ideal leakage for an instance of approximate nearest neighbor search as instantiated in Functionality 1 is captured by $O(d + \log N)$ bits of information per query, where d is the intrinsic dimensionality of the vector space and N is the number of vectors in the database.

Proof. We start by considering the ℓ_∞-norm and induced distance metric Δ. The ℓ_∞-norm is the absolute value of the maximum coordinate: $\ell_\infty||x|| = \max_{\boldsymbol{x}_i} |\boldsymbol{x}_i|$. The induced distance metric is $\Delta(\boldsymbol{x}, \boldsymbol{y}) = \ell_\infty||\boldsymbol{x} - \boldsymbol{y}||$. We first prove the ideal leakage for the ℓ_∞-norm, as it is easier to intuit. We then show how to extend the proof to any ℓ_p-norm induced metric, which includes Euclidean distance. Other common distance metrics, such as angular distance and hamming distance, can be embedded into Euclidean space [36] [17].

Recall that for any query $\boldsymbol{q}$ that will return an ID corresponding to vector $\boldsymbol{v}$, we have that $\Delta(\boldsymbol{v}, \boldsymbol{q}) \leq R_{\mathsf{max}} < D_{\mathsf{max}}$. The ideal functionality implicitly reveals that there exists a point $\boldsymbol{v}$ such that $\Delta(\boldsymbol{v}, \boldsymbol{q}) \leq R_{\mathsf{max}}$. In the ℓ_∞ metric, this implies that $\boldsymbol{v}$ is within a cube of side length $2R_{\mathsf{max}}$ centered at $\boldsymbol{q}$. The output implicitly reveals that $\boldsymbol{v}$ is in this cube. Considering each coordinate of $\boldsymbol{v}$, the number of possible values each coordinate can take diminishes from $2D_{\mathsf{max}}$ to $2R_{\mathsf{max}}$. As a

result, the information revealed is $\log(D_{\max}/R_{\max})$ bits. With the ℓ_∞ metric, this argument can be repeated individually for each coordinate, making the total leakage $d\log(D_{\max}/R_{\max})$ bits.

For any ℓ_p-norm, the above argument holds by considering similar shapes and their relative volumes. In the Euclidean metric $\Delta(\boldsymbol{v},\boldsymbol{q}) \le D_{\max}$ defines a ball of radius $D_{\max}$. The ideal functionality reveals that $\boldsymbol{v}$ is in a ball of radius $R_{\max}$ centered at $\boldsymbol{q}$. The ratio of the volumes of these balls is $D^d_{\max}/R^d_{\max}$, making the leakage $\log(D^d_{\max}/R^d_{\max}) = d\log(D_{\max}/R_{\max})$ bits, as in the ℓ_∞ metric. ∎

Intuitively, this leakage corresponds to the fact that the query vector serves as an approximation for the feature vector of the nearest neighbor, by definition of the problem. The precision of this approximation is limited to $\log(D_{\max}/R_{\max})$.

We now formalize the leakage per query in Protocol 1 and show that it is no worse than $O(d+\log N)$ bits, which matches the leakage of the ideal functionality up to a constant factor.

Claim 3 (Asymptotic Leakage of Protocol 1)**.** Let L be the number of hash tables used to instantiate Protocol 1. Then, the leakage of Protocol 1 is $O(d+\log N)$ bits, matching the ideal functionality.

Proof. Consider the leakage of an infinite number of queries answered through Protocol 1. Fix any $\boldsymbol{v_j} \in \mathcal{DB}$. The total information revealed on $\boldsymbol{v_j}$ is never more than the set $S_j = \{l_1,\ldots,l_L \mid l_i = \mathsf{h}_i(\boldsymbol{v_j})\}$. That is, the set of all L LSH digests of $\boldsymbol{v_j}$. The maximum information revealed is thus the set S_j for all N feature vectors $\boldsymbol{v_j} \in \mathcal{DB}$.

For an individual query, the oblivious masking transformation (Section 5.2) guarantees that at most one element of S—a hash corresponding to a feature vector $\boldsymbol{v} \in \mathcal{DB}$—is leaked. Consider the worst-case (least-hiding) LSH function possible. This is not even a locality-sensitive hash function but instead simply the identity function $\mathsf{h}(\boldsymbol{x}) = \boldsymbol{x}$. It follows that the worst-case implicit query leakage is then $O(d)$ bits, because $l_i = \boldsymbol{v_j}$ for all $i \in \{1,\ldots,L\}$). This is in addition to the explicit $\log N$ bits revealed by the vector ID. Therefore, the total leakage is $O(d+\log N)$, which matches the leakage of the ideal functionality in Theorem 1. ∎

Claim 4. A client deviating from Protocol 1 cannot learn more information on $\mathcal{DB}$ than an honest client following protocol.

Proof. Fix S_j (the set of all hashes) as defined in the proof of Claim 3. By the guarantees of the oblivious masking transformation (Claim 1), even a malicious client can only obtain one element of S_j (for some $j \in \{1,\ldots,L\}$). All other elements are uniformly random or zero. Even if the client does not follow the protocol (such as sending random hashes that don't correspond to any one query), the leakage is capped at a single element of S_j. ∎

In our asymptotic leakage analysis we assume a worst-case scenario in terms of the data and the LSH instantiation. In Appendix B, we analyze the concrete leakage on real data and using a specific instantiation of LSH for Euclidean distance.

8 Empirical evaluation

We now turn to describing our implementation and empirical evaluation of Protocol 1. The goal of this section is to answer the following questions:

- What are the parameters needed to obtain high accuracy in practice using the data structure of Figure 5?
- What is the concrete performance of Protocol 1 when used for ANN search on real data?
- How does Protocol 1 compare to the state-of-the-art approach for private similarity search?

8.1 Implementation and environment

We implement Protocol 1 in approximately 4,000 lines of code. Our implementation is written in Go v1.16. with performance-critical components written in C. The code is open source and available online [2]. Our DPF implementation follows Boyle et al. [30] and is partially based on open-source libraries [45, 48]. Our implementation uses AES as a pseudo-random generator which exploits the AES-NI instruction for hardware-accelerated operations. We use the GMP library [50] for fast modular arithmetic.

8.1.1 Environment

We deploy our implementation on Amazon Elastic Cloud Compute (EC2) for our experiments. We geographically locate the servers on the east coast of the United States while the client is located on the west coast. We measure average ping time of 79 ms between the client and servers. Each server runs on a `c5.9xlarge` Amazon Linux virtual machine. Each server is equipped with Intel Xeon Platinum 8000 CPU (36 vCPUs) and 72 GB of RAM. We use a small `t2.micro` (1 vCPU; 1 GB of RAM) VM for the client (specifications comparable to a low-end smartphone).

8.1.2 Implementation of the LSH

We use the Leech lattice LSH of Andoni and Indyk [8], which we describe in Appendix A for completeness. We find the closest Leech lattice point to a specified point (e.g., the query) using the decoder described by Conway and Sloane [40]. The coordinates of the lattice point are then mapped to the DPF domain size (we set $D = 2^{64}$; Definition 2) using a universal hash. Since the Leech lattice is a 24-dimensional object, the first step of Andoni and Indyk [8, Appendix B] is dimensionality reduction [69]. We randomly project d/k of the coordinates onto each lattice before concatenating the hashes

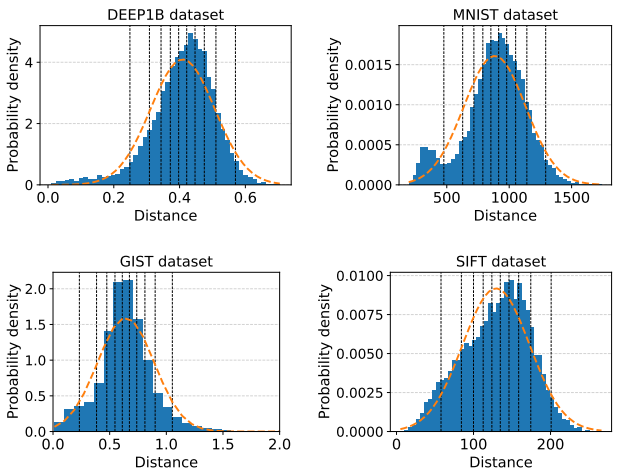

Figure 7: Distances to the nearest neighbor over 10,000 training points in each dataset. The orange curve plots the normal approximation that informs the choice of each R_i for the radix bucketing (see Section 4.2.2). The vertical dashed lines represent the choice of each R_i, for $L = 10$ hash tables.

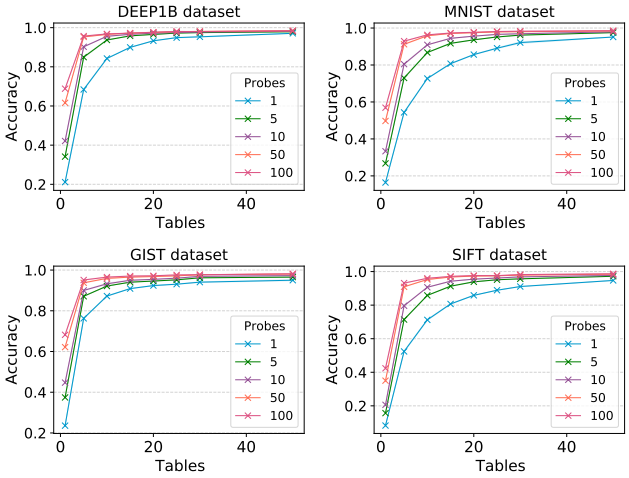

Figure 8: Recall (fraction of approximate nearest neighbors found) for different numbers of table multi-probes using the data structure of Figure 5 and $c = 2$. Average of 10 trial runs for different numbers of probes. Probes = 1 corresponds to no multi-probing (only retrieving the hash of the query from the table; Section 6.1). 95% confidence interval is of size 0.017 and invisible in the plot.

(we use $k = 2$; see Proposition 1). Many locality sensitive hashing algorithms have efficient methods for multi-probing [82]. For lattice based hashes, we choose the multi-probes from the set of closest lattice points, as these correspond to unique hash values. Indeed, the client can advantageously select the closer lattice points when retrieving candidates through PBR (Section 6.2), letting the retrieval failures correspond to probes that are further away. This results in a small increase in multi-probe accuracy.

8.2 Datasets and Parameters

We evaluate our implementation on four real-world datasets obtained from the ANN benchmarks repository [1, 17]. See Table 3 for summary of dataset characteristics.

Dataset	Metric	N	d
DEEP1B	Angular	9,990,000	96
MNIST	Euclidean	60,000	784
GIST	Euclidean	1,000,000	960
SIFT	Euclidean	1,000,000	128

Table 3: Summary of dataset characteristics used in our evaluation.

DEEP1B (Deep billion-scale indexing) [20] contains one billion image embeddings (feature vectors) resulting from the last fully-connected layer of the pretrained GoogLeNet [107] model. We use the smaller version of DEEP1B containing 10M embeddings. Of these, 9,990,000 are used for training (i.e., building the data structure) and 10,000 are used as test queries [17].

MNIST (Modified National Institute of Standards and Technology dataset) [76] contains feature vector representations of handwritten digits. MNIST is widely used to benchmark ANN search algorithms [9, 11, 17, 22]. Each feature vector is a 784 dimensional representation of a 28×28 black and white image of a handwritten digit between zero and nine. MNIST contains 60,000 training feature vectors and 10,000 test queries.

SIFT and GIST [80, 92] are both datasets of image descriptors. Both datasets contains 1,000,000 training feature vectors and 10,000 test queries. In SIFT, each feature vector is 128-dimensional image embedding. In GIST, each feature vector is a 960-dimensional embedding. Both datasets are widely used in ANN algorithm benchmarks [11, 17, 108].

8.2.1 Determining LSH radii

We discuss how to determine the radii to use when instantiating the data structure of Figure 5. We sample 10,000 points from each training set, and find the nearest neighbor among the other training set points. The results are shown in Figure 7. We use a normal distribution computed over the resulting distances as an approximation (orange line in Figure 7), and choose the R_i along the quantiles of the normal distribution. Figure 7 is shown for $L = 10$. This ensures that the number of expected candidates colliding at each R_i is approximately the same (in Figure 7, the area between the dashed lines is the same). Our intuition is that equal areas means that each table will have $\frac{1}{L}$ of the collisions, balancing the load and maximizing efficiency. Each vertical dotted line in Figure 7 corresponds to the R_i for an LSH family for the ith hash table, where $R_1 > D_{\text{min}}$ and $R_{10} = R_{\text{max}}$.

8.3 Performance Evaluation

In this section we report our accuracy and runtime performance. We compare our results to the state-of-the-art approach for privacy-preserving ANN search.

8.3.1 Protocol accuracy and latency

Accuracy. We report the accuracy of Protocol 1 when evaluated on real-world datasets in Figure 8. The datasets we choose form a standard for benchmarking ANN search [17]. We use $k = 2$ for the amplification factor (see Proposition 1). Accuracy is defined in terms of *recall*: what fraction of approximate nearest neighbors found are at most c-times the distance to the true nearest neighbor [6, 55, 63]. Matching the theory, increasing the number of tables *or* multi-probes increases the accuracy. For all datasets we can achieve high accuracy ($> 95\%$) without needing more than 10 tables and 50 multi-probes per table.

Parallelism. The server overhead of answering queries (which involves a linear scan over all hash table keys; see Section 5.4) is easily parallelizable across cores or even across different machines composing each logical server. In our runtime experiments (Figure 9), we provide results for both single core and parallelized executions (where each hash table is processed on a separate core). In our experiments, we observe a linear speedup in the degree of parallelism.

Performance. We measure and report the server-side processing time, communication overhead, and total end-to-end latency to answer each ANN query on a dataset, as measured on the client. The server processing time, per hash table, ranges from 108 ms on the MNIST dataset (60,000 items) to approximately 14 s on the DEEP1B dataset (10,000,000 items). The processing time is dominated by the DPF. The other steps, including ObliviousMasking, take less than 2 ms. The client processing overhead across all datasets was minimal: never exceeding 10 µs. We report the end-to-end latency (as measured on the client machine) for each dataset in Figure 9. We note that the latency includes both the processing time of each server *and* the network delay between the client (located in the `us-west-2` region) and the servers (located in the `us-east-2` region).

Communication. We provide the total communication required per query in Table 4. The communication overhead is determined by three factors: (1) the size of the DPF keys, (2) the number of hash tables L, and (3) the number of multi-probes performed per table (recall Section 6.2). The size of each DPF key is fixed, given that each key encodes a point function in the range $[0, 2^{64}]$ of table bucket hashes (see Section 5.1). As such, only the number of tables L and the number of probes influence the total communication in practice. We report the communication overhead for one table as a function of the number of multi-probes in Table 4.

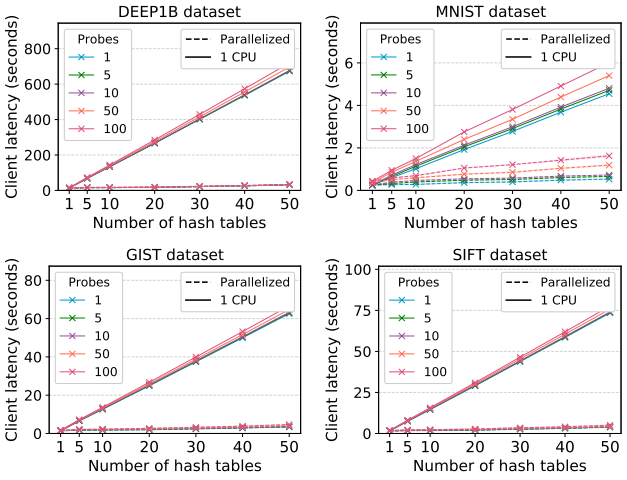

Figure 9: Query latency (including server processing time and network delay) as a function of the number of tables and number of multi-probes performed per table. Probes do not increase computation time but do increase communication resulting in increased latency (see our PBR scheme; Section 6.2). Server parallelization factor is set to equal the number of hash tables. 95% confidence interval is approximately 500 ms and invisible in the plot.

Multi-probes:	**1**	**5**	**10**	**50**	**100**
Communication:	4 kB	13 kB	26 kB	123 kB	245 kB

Table 4: Communication between the client and both servers per hash table in terms of the number of multi-probes performed.

8.3.2 Comparison to SANNS

We compare our approach to SANNS [36]. We note that Chen et al. [36] only evaluate their approach on the DEEP1B and SIFT datasets, and also use the smaller 10M feature vector version of DEEP1B in their evaluation. We use these datasets for a direct comparison in Section 8.3.2. To the best of our knowledge, SANNS is the most efficient privacy-preserving protocol for ANN search, achieving performance on the order of several seconds when evaluated on both DEEP1B and SIFT (over high-bandwidth network connections). To match the evaluation of SANNS, we compare with two network settings. We note that the network configuration used by SANNS has throughput that is *faster* than what we were able to measure on `localhost` using `iperf3` [60], which capped at 3.6 GB/s. The first setting has network throughput between 40 MB/s to 2.2 GB/s, which we call the "fast" network. The second setting has network throughput between 500 MB/s to 7 GB/s, which we call the "`localhost`" network.[5] Given these network configurations, SANNS is by no means deployable over realistic network connections [106], which are over 30× slower. Because SANNS does not provide an open-source implementation, we use the query times reported in their evaluation and note that our deployment environment resembles

[5]Chen et al. [36] refer to this as the "fast" network in their evaluation.

	Protocol 1	SANNS (localhost)	SANNS (Fast Network)
SIFT dataset			
Latency (1 CPU):	70.14 s	8.06 s (8.6× ↓)	59.7 s (1.2× ↓)
Latency (32 CPUs):	2.3 s	1.55 s (1.5× ↓)	14.2 s (6.17× ↑)
Communication:	3.66 MB	1.77 GB (484× ↑)	
DEEP1B dataset			
Latency (1 CPU):	624.3 s	30.1 s (21× ↓)	181 s (3.5× ↓)
Latency (32 CPUs):	21.33 s	4.58 s (4.6× ↓)	37.2 s (1.75× ↑)
Communication:	3.66 MB	5.53 GB (1511× ↑)	

Table 5: End-to-end comparison between Protocol 1 and SANNS over a 500 MB/s to 7 GB/s network (localhost) and a fast network (40 MB/s to 2.2 GB/s). We fix $L = 30$ hash tables and 50 multi-probes per table. SANNS is network-dominated and hence parallelizes less favorably compared to Protocol 1 (see [36, Table 2]).

theirs (comparable CPUs, network, and degree of parallelism applied). We report the results of our comparison in Table 5. Our improvements over SANNS are primarily in terms of communication costs. However, we incur a modest latency overhead on high bandwidth networks, a setting that is highly favorable to SANNS. Our latency is between 21× slower to 6× faster in comparison to SANNS. On the other hand, our communication cost is 484–1511× less in comparison. As a result, over slower networks, e.g., average mobile network supporting 12 Mbps [106], SANNS would incur latency ranging between 19 minutes (for SIFT) to an hour (for DEEP1B), just from the network delay. Over such networks, the latency of Protocol 1 is expected to be 160–500× faster in comparison.

Additionally, because bandwidth costs can be upwards of \$0.02 per GB [19], while CPU cost is around \$0.2 *per hour*, our protocol is monetarily cheaper (per query). SANNS costs \$0.04–\$0.11 per query just for bandwidth alone. In contrast, our protocol costs up to \$0.0002 per query (a 200–550× reduction in total cost).

9 Related work

Existing works on privacy-preserving similarity search either use heavy cryptographic tools (e.g., general secure function evaluation instantiated using two-party computation and fully-homomorphic encryption) or provide poor privacy guarantees for either the client or the database.

To the best of our knowledge, all existing works on privacy-preserving similarity search (with the exception of [104, 119] which use fully-homomorphic encryption) only consider honest-but-curious clients and servers and require many rounds of communication. While generic techniques for upgrading to active security in two-party computation exist [79], they are computationally expensive and often considered impractical. Our protocol is the first to assume fully malicious clients, which solves the challenge left open by prior approaches [36]. A comparison of related work is provided in Table 6.

	Comm.	Comp.	Rounds	Tools	Efficiency
SFR20 [104]	$\log N$	N	1	FHE	●○○○[1]
SANNS [36]	N/k	N	N	2PC	●●○○
IW06 [65]	$\sqrt{N}$	N	N	2PC	N/A[2]
QA08 [95]	N	N	N	2PC	●●○○
ZS21 [119]	$\log N$	N^2	1	FHE	●●○○[3]
Protocol 1	$\sqrt{N}$	N	1	DPF	●●●●

Table 6: Comparison to related-work providing strong privacy guarantees. N is the database size. Ignores constant and $\log N$ factors. We assume dimension $d \in \mathsf{polylog}(N)$ [55].

[1] Uses fully-homomorphic encryption to evaluate the entire computation; takes several hours of computation to evaluate on very small datasets.
[2] Theoretical work; no implementation or concrete runtime estimates provided.
[3] Uses fully-homomorphic encryption to evaluate the entire computation; takes over 20 minutes to evaluate on very small datasets ($N \approx 1000$).

Two-party computation based approaches. Indyk and Woodruff [65] investigate nearest neighbor search between two parties under the Euclidean distance metric. They show a $\tilde{O}(\sqrt{N})$ communication protocol for finding an approximate *near* (as opposed to *nearest*) neighbor to a query. Their techniques rely on black-box two-party computation. This makes them only asymptotically efficient (they do not provide an implementation or any concrete efficiency estimate). However, their protocol shares some similarity to ours. Specifically, they tolerate some precisely quantified leakage, which they argue can be a suitable compromise in favor of efficiency gains.

More recently, Chen et al. [36] design and evaluate SANNS, a system for approximate nearest neighbor search that uses oblivious RAM, garbled circuits, and homomorphic encryption. Their solution combines heuristic k-means clustering techniques to reduce overhead of two-party computation of computing oblivious comparisons by a constant factor (i.e., $\frac{1}{k}$). However, they still require asymptotically linear communication, since k is typically small. They leave open the possibility of using locality-sensitive hashing to provide *provable* guarantees, as we do.

By assuming data-dependent clusters, SANNS is able to forgo a linear scan over the database when evaluating oblivious distance comparisons. While their implementation is efficient over high-bandwidth connections (40MB/s to 7GB/s bandwidth), such throughput is often too high to be practical on real-world connections (especially on mobile networks, which can be over 30× slower [106]). Moreover, SANNS requires the data to be clusterable to reduce communication. Our protocol, in contrast, does not make any assumptions on the input data to achieve sublinear communication. We show in our evaluation (Section 8) that our protocol requires between 484–1511× less communication compared to SANNS, and microseconds of processing on the client.

Qi and Atallah [95] present a protocol for privacy-preserving

nearest neighbor search in the honest-but-curious setting with two parties. In contrast to us, they assume each party (i.e., server) has a database that is private from the other party. Queries are computed over the union of both databases. Their protocol uses secure two-party computation to compute oblivious comparisons and requires linear communication in the database size. Qi and Atallah [95] do not provide an implementation or any concrete efficiency estimates for their protocol.

Fully-homomorphic encryption based approaches. Shaul et al. [104] present a protocol based on fully-homomorphic encryption, requiring several hours of computation time to answer queries over small (1000 item) databases. While this results in both a single-round protocol and tolerates malicious clients, it is not practical for large databases. Their implementation requires between three and eight hours (parallelized across 16 cores) to compute the nearest neighbors on small datasets ranging between 1,000 and 4,000 feature vectors.

Zuber and Sirdey [119] implement a secure nearest neighbor classifier using (threshold) fully-homomorphic encryption with applications to collaborative learning and nearest neighbor search. Their approach requires over one hour of server processing time to compute a query answer over a small database of approximately 500 feature vectors. As such, their protocol is not scalable beyond databases containing a few thousand feature vectors.

Partially-private approaches. Not directly related to privacy, Aumüller et al. [18] introduce distance-sensitive hashing which they show can be beneficial to reducing information leakage between hashes. However, their security guarantees are not formally defined and their approach provides a trade-off between privacy and accuracy, leaking information about the client's query and the database simultaneously.

Riazi et al. [101] likewise explore LSH as a means of trading-off privacy with accuracy, with more accurate results revealing more information on both the query and the database. They make use of two-party computation to instantiate a garbled circuit for the purpose of securely evaluating locality-sensitive hashes without revealing the description of the hash function. While this reduces some leakage, their approach still reveals information on the query and the database.

In a similar vein, Boufounos and Rane [28] develop a binary embedding (locality-sensitive hash) that preserves privacy when finding similar feature vectors in a remote database. Their technique is less general compared to [18, 101]. Boufounos and Rane [28] do not provide a formal security analysis of their nearest neighbor search protocol based on their secure embedding. Analyzing their protocol, we found that (1) some partial information on the query is inadvertently leaked to the server, and (2) the client learns the distances from its query to all near-neighbors, resulting in significant database leakage. No implementation or any concrete runtime estimates were provided for their protocol.

Client-outsourced data. Other systems for privacy-preserving nearest neighbor search (e.g., [15, 47, 62, 75, 114]) focus on a setting where a client *outsources* the data to a remote database in a way that it can later privately query for nearest neighbors. As such, they do not consider malicious clients nor are they interested in preserving the privacy of item attributes in the database. We consider this setting as orthogonal.

10 Discussion

Using two servers to instantiate lightweight privacy-preserving systems has proven fruitful for a wide range of applications [5, 27, 37, 42, 43, 45, 46, 48, 73, 88], including systems deployed in industry [45, 59]. In this paper, we presented the first lightweight privacy-preserving system for approximate nearest neighbor search using the two-server model. The resulting protocol is highly efficient, with only one round and a few megabytes of communication between the client and servers. Moreover, we do not require servers to communicate when answering queries. This is in contrast to many previous uses of the two-server model, which typically also require communication between servers when processing requests. Our evaluation (Section 8) provides evidence for the real-world practicality of our approach, since it is deployable on high-latency, low-bandwidth networks with lightweight clients (e.g., low-end smartphones). We survey some potential applications of our protocol.

Use case 1: Private (approximate) DNA matching. DNA sequencing has gained significant popularity in the last decade, leading to swaths of genome data with a wide array of applications (see Naveed et al. [87] for a survey of applications). Such data, while highly valuable, is also highly personal and can uniquely identify individuals [68, 87]. One application of DNA sequence matching, highlighted by Naveed et al. [87, Section 7.1] is personalized healthcare. For example, a patient may be interested in finding whether or not they are prone to a certain disease. Patients submit their genome sequence (query) to an organization (e.g., a pharmaceutical company) mapping sequences to various disease markers (IDs). When queried, the organization finds common disease markers associated with the supplied genome sequence. In this application, both the patient and the organization require privacy. The patient would like to keep their genome sequence private while the organization requires privacy for their potentially proprietary sequences associated with various disease markers [26, 51, 110]. The pattern matching can be viewed as a nearest neighbor problem. Indeed, Buhler [31] shows how to apply locality-sensitive hashing to approximate DNA sequence matching. As such, our protocol may be suitable to this application, enabling privacy-preserving services for personalized medicine.

Use case 2: Private targeted advertising. Online targeted

advertising has been widely criticized for the rampant privacy violations it incurs [58, 90, 105]. Companies have been found to use protected demographics like gender, race, or religion to target advertisements in a discriminatory manner [7, 33, 64, 113]. Indeed, recent regulations, introduced in Europe and California, forbid the collection and use of user data for the purpose of advertising without explicit consent [3, 4]. Several privacy-preserving systems for targeted advertising have been proposed to address these issues [21, 61, 70, 94, 102, 109, 112]. The core component of these systems is a privacy-preserving *targeting* mechanism which matches users with relevant ads. This targeting process can be based on LSH [102, 112]. For example, in AdVeil [102], single-server PIR is used to query a targeting data structure. While user privacy is maintained, AdVeil does not take database privacy into consideration (which may contain proprietary targeting data). Similarly, with FLoC [112], LSH is used to group users into "cohorts" for which relevant ads are served. The smaller the cohort, the less privacy users can expect from the system, resulting in an undesirable privacy-accuracy trade-off in the ad targeting mechanism. Both these systems could potentially be extended with our approach to similarity search for the purpose of introducing better database and user privacy.

Use case 3: Private medical diagnoses. Zuber and Sirdey [119] highlight a use case involving classifying images for medical diagnoses. They show how their nearest neighbor classifier can be used to (privately) classify tumors as either malignant or benign using the Breast Cancer Wisconsin diagnostic dataset [16]. Zuber and Sirdey [119] perform this classification using fully-homomorphic encryption, (see Section 9). Our protocol could be applied to this setting, with very minimal overhead. For example, the Breast Cancer dataset contains 569 feature vectors. We can assign the ID of each feature vector to be either 1 or 2, for malignant or benign. Finding the nearest neighbor in this dataset can be performed using Protocol 1 in under 10 ms of end-to-end latency. Even on much larger datasets (e.g., MNIST with 60,000 feature vectors) latency is under 900ms per query, as was shown in Section 8.

11 Conclusion

We presented a new protocol for privacy-preserving similarity search based on locality-sensitive hashing. Our protocol is lightweight, incurring almost no processing overhead on the client while remaining concretely efficient for the database servers. We showed that clients can obtain answers to queries within a few seconds of latency, even on slow networks and large databases. Compared to the state-of-the-art in privacy-preserving approximate neighbor search, which requires many rounds and *gigabytes* of communication, our protocol requires only one round and less than four *megabytes* of communication. We achieve these gains by tolerating some extra database leakage to the client (while maintaining complete client privacy). However, we carefully analyze and bound the database leakage, which we show is asymptotically optimal (matching the ideal functionality) but concretely reveals a bit more information on the database. We show that this leakage is bounded, even in the face of malicious clients deviating from protocol. Finally, we surveyed several applications of our protocol, which range from privacy-preserving personalized medicine to targeted advertising.

12 Acknowledgements

We thank Kyle Hogan, Zack Newman, and Leah Rosenbloom for helpful feedback on early drafts of this paper.

References

[1] ANN benchmark datasets. https://github.com/erikbern/ann-benchmarks/. Accessed August 2021.

[2] Source code. https://github.com/sachaservan/private-ann.

[3] General Data Protection Regulation. https://gdpr-info.eu/, 2016. Accessed August 2021.

[4] California Consumer Privacy Act of 2018. https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5, 2018. Accessed August 2021.

[5] Surya Addanki, Kevin Garbe, Eli Jaffe, Rafail Ostrovsky, and Antigoni Polychroniadou. Prio+: Privacy preserving aggregate statistics via boolean shares.

[6] Thomas D Ahle, Martin Aumüller, and Rasmus Pagh. Parameter-free locality sensitive hashing for spherical range reporting. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 239–256. SIAM, 2017.

[7] Muhammad Ali, Piotr Sapiezynski, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Ad delivery algorithms: The hidden arbiters of political messaging. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 13–21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382977. doi: 10.1145/3437963.3441801. URL https://doi.org/10.1145/3437963.3441801.

[8] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 459–468, 2006. doi: 10.1109/FOCS.2006.49.

[9] Alexandr Andoni, Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab Mirrokni. Locality-sensitive hashing using stable distributions. *Nearest-Neighbor Methods in Learning and Vision, ed. G. Shakhnarovich, T. Darrel, and P. Indyk*, pages 61–72, 2005.

[10] Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. Practical and optimal lsh for angular distance. *arXiv preprint arXiv:1509.02897*, 2015.

[11] Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya P. Razenshteyn, and Ludwig Schmidt. Practical and optimal LSH for angular distance. *CoRR*, abs/1509.02897, 2015. URL http://arxiv.org/abs/1509.02897.

[12] Alexandr Andoni, Piotr Indyk, and Ilya Razenshteyn. Approximate nearest neighbor search in high dimensions. *arXiv preprint arXiv:1806.09823*, 7, 2018.

[13] Sebastian Angel and Srinath Setty. Unobservable communication over fully untrusted infrastructure. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 551–569, 2016.

[14] Sebastian Angel, Hao Chen, Kim Laine, and Srinath Setty. Pir with compressed queries and amortized query processing. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 962–979. IEEE, 2018.

[15] Gilad Asharov, Shai Halevi, Yehuda Lindell, and Tal Rabin. Privacy-preserving search of similar patients in genomic data. *Proc. Priv. Enhancing Technol.*, 2018 (4):104–124, 2018.

[16] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[17] Martin Aumüller, Erik Bernhardsson, and Alexander John Faithfull. Ann-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *CoRR*, abs/1807.05614, 2018. URL http://arxiv.org/abs/1807.05614.

[18] Martin Aumüller, Tobias Christiani, Rasmus Pagh, and Francesco Silvestri. Distance-sensitive hashing. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 89–104, 2018.

[19] Azure. Microsoft Azure bandwidth pricing. https://azure.microsoft.com/pricing/details/bandwidth/. Accessed August 2021.

[20] Artem Babenko and Victor Lempitsky. Efficient indexing of billion-scale datasets of deep descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2055–2063, 2016.

[21] Michael Backes, Aniket Kate, Matteo Maffei, and Kim Pecina. Obliviad: Provably secure and practical online behavioral advertising. In *2012 IEEE Symposium on Security and Privacy*, pages 257–271. IEEE, 2012.

[22] Arturs Backurs, Yihe Dong, Piotr Indyk, Ilya Razenshteyn, and Tal Wagner. Scalable nearest neighbor search for optimal transport. In *International Conference on Machine Learning*, pages 497–506. PMLR, 2020.

[23] Mauro Barni, Tiziano Bianchi, Dario Catalano, Mario Di Raimondo, Ruggero Donida Labati, Pierluigi Failla, Dario Fiore, Riccardo Lazzeretti, Vincenzo Piuri, Fabio Scotti, et al. Privacy-preserving fingercode authentication. In *Proceedings of the 12th ACM workshop on Multimedia and security*, pages 231–240, 2010.

[24] Michael Ben-Or, Shafi Goldwasser, and Avi Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*, pages 351–371. 2019.

[25] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999.

[26] Marina Blanton, Mikhail J Atallah, Keith B Frikken, and Qutaibah Malluhi. Secure and efficient outsourcing of sequence comparisons. In *European Symposium on Research in Computer Security*, pages 505–522. Springer, 2012.

[27] Dan Boneh, Elette Boyle, Henry Corrigan-Gibbs, Niv Gilboa, and Yuval Ishai. Lightweight techniques for private heavy hitters. *arXiv preprint arXiv:2012.14884*, 2020.

[28] Petros Boufounos and Shantanu Rane. Secure binary embeddings for privacy preserving nearest neighbors. In *2011 IEEE International Workshop on Information Forensics and Security*, pages 1–6. IEEE, 2011.

[29] Elette Boyle, Niv Gilboa, and Yuval Ishai. Function secret sharing. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 337–367. Springer, 2015.

[30] Elette Boyle, Niv Gilboa, and Yuval Ishai. Function secret sharing: Improvements and extensions. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1292–1303, 2016.

[31] Jeremy Buhler. Efficient large-scale sequence comparison by locality-sensitive hashing. *Bioinformatics*, 17 (5):419–428, 2001.

[32] Paul Bunn, Jonathan Katz, Eyal Kushilevitz, and Rafail Ostrovsky. Efficient 3-party distributed oram. In *International Conference on Security and Cryptography for Networks*, pages 215–232. Springer, 2020.

[33] José González Cabañas, Ángel Cuevas, and Rubén Cuevas. Unveiling and quantifying Facebook exploitation of sensitive personal data for advertising purposes. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 479–495, Baltimore, MD, August 2018. USENIX Association. ISBN 978-1-939133-04-5.

[34] Rui Cai, Chao Zhang, Lei Zhang, and Wei-Ying Ma. Scalable music recommendation by search. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 1065–1074, 2007.

[35] Xia Cao, Shuai Cheng Li, Beng Chin Ooi, and Anthony KH Tung. Piers: An efficient model for similarity search in dna sequence databases. *ACM Sigmod Record*, 33(2):39–44, 2004.

[36] Hao Chen, Ilaria Chillotti, Yihe Dong, Oxana Poburinnaya, Ilya Razenshteyn, and M Sadegh Riazi. SANNS: Scaling up secure approximate k-nearest neighbors search. In *29th USENIX Security Symposium (USENIX Security 20)*, 2020.

[37] Weikeng Chen and Raluca Ada Popa. Metal: A metadata-hiding file-sharing system. *IACR Cryptol. ePrint Arch.*, 2020:83, 2020.

[38] Benny Chor, Oded Goldreich, Eyal Kushilevitz, and Madhu Sudan. Private information retrieval. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 41–50. IEEE, 1995.

[39] Benny Chor, Niv Gilboa, and Moni Naor. *Private information retrieval by keywords*. Citeseer, 1997.

[40] J. Conway and N. Sloane. Soft decoding techniques for codes and lattices, including the golay code and the leech lattice. *IEEE Transactions on Information Theory*, 32(1):41–50, 1986. doi: 10.1109/TIT.1986.1057135.

[41] J. H. Conway, R. A. Parker, and N. J. A. Sloane. The covering radius of the leech lattice. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 380(1779):261–290, 1982. ISSN 00804630. URL http://www.jstor.org/stable/2397303.

[42] Henry Corrigan-Gibbs and Dan Boneh. Prio: Private, robust, and scalable computation of aggregate statistics. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, pages 259–282, 2017.

[43] Henry Corrigan-Gibbs, Dan Boneh, and David Mazières. Riposte: An anonymous messaging system handling millions of users. In *2015 IEEE Symposium on Security and Privacy*, pages 321–338. IEEE, 2015.

[44] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262, 2004.

[45] Emma Dauterman, Eric Feng, Ellen Luo, Raluca Ada Popa, and Ion Stoica. DORY: An encrypted search system with distributed trust. In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, pages 1101–1119, 2020.

[46] Jack Doerner and Abhi Shelat. Scaling oram for secure computation. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 523–535, 2017.

[47] Yousef Elmehdwi, Bharath K Samanthula, and Wei Jiang. Secure k-nearest neighbor query over encrypted data in outsourced environments. In *2014 IEEE 30th International Conference on Data Engineering*, pages 664–675. IEEE, 2014.

[48] Saba Eskandarian, Henry Corrigan-Gibbs, Matei Zaharia, and Dan Boneh. Express: Lowering the cost of metadata-hiding communication with cryptographic privacy. *arXiv preprint arXiv:1911.09215*, 2019.

[49] David Evans, Yan Huang, Jonathan Katz, and Lior Malka. Efficient privacy-preserving biometric identification. In *Proceedings of the 17th conference Network and Distributed System Security Symposium, NDSS*, volume 68, pages 90–98, 2011.

[50] GNU foundation. GMP library. https://gmplib.org/. Accessed August 2021.

[51] Keith B Frikken. Practical private dna string searching and matching through efficient oblivious automata evaluation. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 81–94. Springer, 2009.

[52] William Gasarch. A survey on private information retrieval. *Bulletin of the EATCS*, 82(72-107):113, 2004.

[53] Yael Gertner, Yuval Ishai, Eyal Kushilevitz, and Tal Malkin. Protecting data privacy in private information retrieval schemes. *Journal of Computer and System Sciences*, 60(3):592–629, 2000.

[54] Niv Gilboa and Yuval Ishai. Distributed point functions and their applications. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 640–658. Springer, 2014.

[55] Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via hashing. In *Vldb*, volume 99, pages 518–529, 1999.

[56] Oded Goldreich, Howard Karloff, Leonard J Schulman, and Luca Trevisan. Lower bounds for linear locally decodable codes and private information retrieval. In *Proceedings 17th IEEE Annual Conference on Computational Complexity*, pages 175–183. IEEE, 2002.

[57] Oded Goldreich, Silvio Micali, and Avi Wigderson. How to play any mental game, or a completeness theorem for protocols with honest majority. In *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*, pages 307–328. 2019.

[58] Bill Goodwin and Sebastian Klovig Skelton. Facebook's privacy game – how Zuckerberg backtracked on promises to protect personal data. https://www.computerweekly.com/feature/Facebooks-privacy-U-turn-how-Zuckerberg-backtracked-on-promises-to-protect-personal-data, 2019. Accessed August 2021.

[59] Internet Security Research Group. Prio services: Privacy-respecting application metrics. https://www.abetterinternet.org/prio/, 2021. Accessed August 2021.

[60] Vivien Gueant. iPerf—the TCP, UDP and SCTP network bandwidth measurement tool. 2021. Accessed August 2021.

[61] Saikat Guha, Bin Cheng, and Paul Francis. Privad: Practical privacy in online advertising. In *USENIX conference on Networked systems design and implementation*, pages 169–182, 2011.

[62] Cheng Guo, Shenghao Su, Kim-Kwang Raymond Choo, and Xinyu Tang. A fast nearest neighbor search scheme over outsourced encrypted medical images. *IEEE Transactions on Industrial Informatics*, 17(1):514–523, 2018.

[63] Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. *Theory of computing*, 8(1): 321–350, 2012.

[64] Basileal Imana, Aleksandra Korolova, and John Heidemann. Auditing for discrimination in algorithms delivering job ads. WWW '21, page 3767–3778, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450077. URL https://doi.org/10.1145/3442381.3450077.

[65] Piotr Indyk and David Woodruff. Polylogarithmic private approximations and efficient matching. In *Theory of Cryptography Conference*, pages 245–264. Springer, 2006.

[66] Yuval Ishai, Eyal Kushilevitz, Rafail Ostrovsky, and Amit Sahai. Batch codes and their applications. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 262–271, 2004.

[67] Yuval Ishai, Eyal Kushilevitz, Rafail Ostrovsky, and Amit Sahai. Batch codes and their applications. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 262–271, 2004.

[68] Aaron Johnson and Vitaly Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1079–1087, 2013.

[69] William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space 26. *Contemporary mathematics*, 26, 1984.

[70] Ari Juels. Targeted advertising... and privacy too. In *Cryptographers' Track at the RSA Conference*, pages 408–424. Springer, 2001.

[71] Shakira Banu Kaleel and Abdolreza Abhari. Cluster-discovery of twitter messages for event detection and trending. *Journal of computational science*, 6:47–57, 2015.

[72] Donald Ervin Knuth. *The art of computer programming*, volume 3. Pearson Education, 1997.

[73] Dmitry Kogan and Henry Corrigan-Gibbs. Private blocklist lookups with checklist. *IACR Cryptol. ePrint Arch.*, 2021:345, 2021.

[74] Brian Kulis and Kristen Grauman. Kernelized locality-sensitive hashing for scalable image search. In *2009 IEEE 12th international conference on computer vision*, pages 2130–2137. IEEE, 2009.

[75] Mehmet Kuzu, Mohammad Saiful Islam, and Murat Kantarcioglu. Efficient similarity search over encrypted data. In *2012 IEEE 28th International Conference on Data Engineering*, pages 1156–1167. IEEE, 2012.

[76] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[77] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[78] Yury Lifshits. The homepage of nearest neighbors and similarity search. http://simsearch.yury.name/, 2008. Accessed August 2021.

[79] Yehuda Lindell and Benny Pinkas. An efficient protocol for secure two-party computation in the presence of malicious adversaries. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 52–78. Springer, 2007.

[80] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[81] Dandan Lu, Yue Zhang, Ling Zhang, Haiyan Wang, Wanlin Weng, Li Li, and Hongmin Cai. Methods of privacy-preserving genomic sequencing data alignments. *Briefings in Bioinformatics*, 2021.

[82] Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li. Multi-probe lsh: efficient indexing for high-dimensional similarity search. In *33rd International Conference on Very Large Data Bases, VLDB 2007*, pages 950–961. Association for Computing Machinery, Inc, 2007.

[83] Guillaume Marçais, Dan DeBlasio, Prashant Pandey, and Carl Kingsford. Locality-sensitive hashing for the edit distance. *Bioinformatics*, 35(14):i127–i135, 2019.

[84] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*, 2013.

[85] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

[86] Rajeev Motwani, Assaf Naor, and Rina Panigrahi. Lower bounds on locality sensitive hashing. In *Proceedings of the twenty-second annual symposium on Computational geometry*, pages 154–157, 2006.

[87] Muhammad Naveed, Erman Ayday, Ellen W Clayton, Jacques Fellay, Carl A Gunter, Jean-Pierre Hubaux, Bradley A Malin, and XiaoFeng Wang. Privacy in the genomic era. *ACM Computing Surveys (CSUR)*, 48(1): 1–44, 2015.

[88] Zachary Newman, Sacha Servan-Schreiber, and Srinivas Devadas. Spectrum: High-bandwidth anonymous broadcast with malicious security. *IACR Cryptol. ePrint Arch.*, 2021:325, 2021.

[89] Laurent Noé and Gregory Kucherov. *YASS: Similarity search in DNA sequences*. PhD thesis, INRIA, 2003.

[90] Kate O'Flaherty. This is why people no longer trust Google and Facebook with their data. https://www.forbes.com/sites/kateoflahertyuk/2018/10/10/this-is-why-people-no-longer-trust-google-and-facebook-with-their-data, 2018. Accessed August 2021.

[91] Naoaki Okazaki and Jun'ichi Tsujii. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 851–859, 2010.

[92] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42 (3):145–175, 2001.

[93] Rina Panigrahy. Entropy based nearest neighbor search in high dimensions. *arXiv preprint cs/0510019*, 2005.

[94] Gonçalo Pestana, Iñigo Querejeta-Azurmendi, Panagiotis Papadopoulos, and Benjamin Livshits. THEMIS: Decentralized and trustless ad platform with reporting integrity, 2020.

[95] Yinian Qi and Mikhail J Atallah. Efficient privacy-preserving k-nearest neighbor search. In *2008 The 28th International Conference on Distributed Computing Systems*, pages 311–319. IEEE, 2008.

[96] Michael O Rabin. How to exchange secrets with oblivious transfer. *IACR Cryptol. ePrint Arch.*, 2005 (187), 2005.

[97] Anand Rajaraman and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

[98] Anand Rajaraman and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

[99] Deepak Ravichandran, Patrick Pantel, and Eduard Hovy. Randomized algorithms and nlp: Using locality sensitive hash functions for high speed noun clustering. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 622–629, 2005.

[100] Eric Rescorla and Tim Dierks. The transport layer security (tls) protocol version 1.3. 2018.

[101] M Sadegh Riazi, Beidi Chen, Anshumali Shrivastava, Dan Wallach, and Farinaz Koushanfar. Sublinear privacy-preserving near-neighbor search. *arXiv preprint arXiv:1612.01835*, 2016.

[102] Sacha Servan-Schreiber, Kyle Hogan, and Srinivas Devadas. Adveil: A private targeted-advertising ecosystem. *IACR Cryptol. ePrint Arch.*, 2021:1032, 2021.

[103] Adi Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

[104] Hayim Shaul, Dan Feldman, and Daniela Rus. Secure k-ish nearest neighbors classifier. *Proceedings on Privacy Enhancing Technologies*, 2020(3):42–61, 2020.

[105] Natasha Singer. What you don't know about how Facebook uses your data. https://www.nytimes.com/2018/04/11/technology/facebook-privacy-hearings.html, 2018.

[106] Speedtest.com. Speedtest.com global index. https://www.speedtest.net/global-index, 2021. Accessed August 2021.

[107] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[108] TEXMEX Research Team. Datasets for approximate nearest neighbor search. http://corpus-texmex.irisa.fr/, 2010. Accessed August 2021.

[109] Vincent Toubiana, Arvind Narayanan, Dan Boneh, Helen Nissenbaum, and Solon Barocas. Adnostic: Privacy preserving targeted advertising. In *Proceedings Network and Distributed System Symposium*, 2010.

[110] Juan Ramón Troncoso-Pastoriza, Stefan Katzenbeisser, and Mehmet Celik. Privacy preserving error resilient dna searching through oblivious automata. In *Proceedings of the 14th ACM conference on Computer and communications security*, pages 519–528, 2007.

[111] Avery Wang et al. An industrial strength audio search algorithm. In *Ismir*, volume 2003, pages 7–13. Citeseer, 2003.

[112] Web Incubator CG. FLoC. https://github.com/WICG/floc, 2021. Accessed August 2021.

[113] Miranda Wei, Madison Stamos, Sophie Veys, Nathan Reitinger, Justin Goodman, Margot Herman, Dorota Filipczuk, Ben Weinshel, Michelle L. Mazurek, and Blase Ur. What Twitter knows: Characterizing ad targeting practices, user perceptions, and ad explanations through users' own twitter data. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 145–162. USENIX Association, August 2020. ISBN 978-1-939133-17-5.

[114] Wei Wu, Udaya Parampalli, Jian Liu, and Ming Xian. Privacy preserving k-nearest neighbor classification over encrypted database in outsourced cloud environments. *World Wide Web*, 22(1):101–123, 2019.

[115] Guowen Xu, Hongwei Li, Hao Ren, Xiaodong Lin, and Xuemin Sherman Shen. Dna similarity search with access control over encrypted cloud data. *IEEE Transactions on Cloud Computing*, 2020.

[116] Andrew C Yao. Protocols for secure computations. In *23rd annual symposium on foundations of computer science (sfcs 1982)*, pages 160–164. IEEE, 1982.

[117] Peter N Yianilos. Locally lifting the curse of dimensionality for nearest neighbor search. 1999.

[118] Clara E Yoon, Ossian O'Reilly, Karianne J Bergen, and Gregory C Beroza. Earthquake detection through computationally efficient similarity search. *Science advances*, 1(11):e1501057, 2015.

[119] Martin Zuber and Renaud Sirdey. Efficient homomorphic evaluation of k-nn classifiers. *Proc. Priv. Enhancing Technol.*, 2021(2):111–129, 2021.

A Efficient LSH for Euclidean distance

In this section we describe how to instantiate Proposition 1 efficiently for Euclidean distance ANN search. We first note that Proposition 1 is an upper bound and assumes worst-case data (all points are between R and cR from the query). In practice, there are LSH functions that can achieve very small p_2 without needing a large k. We take another look at Proposition 1. For a large enough $k > d$, the region of space defined by the set $\{x \mid \mathsf{h}(x) = \mathsf{h}(y)\}$ for a fixed y becomes bounded. Specifically, there exists a bounding distance b such that

$$\Pr[h(x) = h(y) \wedge ||x - y|| > b] < \delta.$$

For a false-positive rate of δ, it suffices to scale the space so that the bounding distance b is equal to cR. This scaling decreases p_1 as explained in the proof of Proposition 1. For better efficiency, we observe that it is possible to use LSH functions that *inherently* have bounded regions, such as the Leech lattice based LSH of Andoni and Indyk [8]. By using structure rather than randomness they can achieve a larger p_1 for the same p_2 as one might expect from $k \approx d$. We briefly describe a lattice-based LSH family for Euclidean distance next.

Lattice-based LSH. A simple lattice-based locality sensitive hash family for Euclidean distance is as follows. Choose an infinite set of points in $\mathbb{R}^d$, and let the hash of x be the closest point of this set. For example, if we consider the set of points with integer coordinates, we can efficiently find the hash by rounding the coordinates of x. Lattice based hashes allow us to bound the error distance on p_2; with the integer points, it is clear that $b \leq \sqrt{d}$. Furthermore, we can reduce b to b' by returning $\perp$ if the distance $||x - h(x)||$ is not less than $b'/2$. However, the set of integer points is not optimal in that the error increases greatly for the higher dimensional data we wish to use it on. For an efficient implementation, we use the Leech lattice (as in [8, Appendix B]), which has $b \leq \sqrt{2}$ for $d = 24$. See Conway et al. [41] or Conway and Sloane [40] for more information on the lattice structure.

B Bounding concrete leakage

From the proof of Claim 3, we see that the asymptotic leakage consists of at most one LSH digest, even when the client is acting maliciously (Claim 4). In the worst case, this digest corresponds to a full feature vector, leading to the asymptotic bound. However, concretely, we would like to analyze *how much worse* Protocol 1 is in comparison to Functionality 1. We answer this in Claim 5. We show that Protocol 1, when instantiated for Euclidean distance using the Leech lattice LSH (Appendix A), leaks a more precise approximation compared to Functionality 1. Intuitively, this extra leakage comes from the client learning which radix bucket the nearest neighbor to the query is located in. To analyze the concrete leakage, we must first establish some technical groundwork pertaining to the Leech-lattice LSH we use for Euclidean (and Angular) distance. We do so in Lemma 1, where we will show that a Leech-lattice LSH digest is inherently less precise than the corresponding ideal ball in Euclidean space. Specifically, we show that the ideal ball of radius R has a smaller volume compared to the region of space represented by an LSH digest.

Lemma 1. *The* $(1, 2, 0.0097459, 0.0000156)$*-sensitive Leech lattice LSH [8] has volume at least* $\frac{V_B}{0.77}$*, where* V_B *is the volume of the* 24*-dimensional unit ball.*

Proof. Our proof relies on the following fact about the Leech lattice.

> **Fact 1:** The Leech lattice is a 24 dimensional object with a sphere packing density of $\frac{\pi^{12}}{12!}$ [41], where the density is defined as the fraction of space covered by (tangent) balls centered at the lattice points.

By Fact 1, we have that the ratio of the volume of the unit ball to that of the lattice cell is given by the density. A small calculation shows that the lattice cell then has volume 1. We ask what is the minimum radius for a 24-dimensional ball such that the ball has volume 1. We find that the answer is

$$\sqrt[24]{\frac{12!}{\pi^{12}}} \approx 1.29.$$

The inverse of this quantity is 0.77, which proves the lemma [6]. The values of $p_1 = 0.0097459$ and $p_2 = 0.0000156$ for $c = 2$ are from [8, Table 1]. ∎

Remark 2. In Lemma 1 we describe the $(1, 2, 0.0097459, 0.0000156)$-sensitive Leech lattice LSH for Euclidean space as described by Andoni and Indyk [8]. We note that this LSH can be used for any R and cR (while also preserving the same p_1 and p_2), by simply scaling the input vectors accordingly. See Andoni and Indyk [8] for a more detailed explanation.

Using Lemma 1, we can bound the concrete leakage of Protocol 1 when instantiated with the Leech-lattice LSH. This allows us to empirically bound the concrete leakage for the datasets we use in our evaluation (Section 8). We compute a precise leakage bound in Claim 5.

Claim 5 (Concrete Leakage of Protocol 1)**.** Fix R_{min} and R_{max} as defined in both Figure 5 and Functionality 1, where $D_{\mathsf{min}} < R_{\mathsf{min}} \leq R_{\mathsf{max}} < D_{\mathsf{max}}$. Protocol 1, when instantiated with the Leech-lattice based LSH of Andoni and Indyk [8] leaks at most a multiplicative factor of $0.77 \cdot \frac{R_{\mathsf{max}}}{R_{\mathsf{min}}}$ more information compared to Functionality 1.

Proof. We recall the argument in the proof of Claim 3. Fix any $\boldsymbol{v_j} \in \mathcal{DB}$ and S_j as in the proof of Claim 3. We show that S_j is an upper bound on what can be revealed on $\boldsymbol{v_j}$ through Protocol 1. We now examine how much faster this set can be leaked with queries issued to Protocol 1 and contrast it to Functionality 1.

Define an ideal oracle $\mathcal{O}(\mathcal{DB}, R_i, \cdot)$, which given a query $\boldsymbol{q}$, outputs the ID of any vector in $\mathcal{DB}$ within distance R_i of $\boldsymbol{q}$, if such a vector exists. Observe that Functionality 1 is modeled by $\mathcal{O}(\mathcal{DB}, R_{\mathsf{max}}, \cdot)$.

Suppose that the client obtains a query answer and learns $l_i \in S_j$. Note that this leakage is less than (or equal to) that of learning $l_1 \in S_j$. That is, the hash of the same vector but on the smallest radius. It is easy to see that l_1 is at least as precise (contains as much information) as $l_i \in S_j$, $i > 1$, due to the increasing radii of the radix buckets, as explained in Section 4.2.1.

Next, by Lemma 1, we have that the information revealed by $l_1 = \mathsf{h}_1(\boldsymbol{q})$ is less than or equal to the information revealed by $\mathcal{O}(\mathcal{DB}, R_1, \boldsymbol{q})$. Indeed, as was shown in Lemma 1, l_i is 0.77 less precise compared to $\mathcal{O}(\mathcal{DB}, R_i, \boldsymbol{q})$, for any i.

Therefore, we have that the ratio in precision between l_1 and $\mathcal{O}(\mathcal{DB}, R_{\mathsf{max}}, \cdot)$, that is, the ratio between the precision of the smallest radix bucket and the ideal functionality, is bounded by 0.77 times the ratio of precision between the ideal oracles

[6] For the $(1.2, 1.8, 0.00515, 0.0000771)$-sensitive LSH [8], this factor is $\frac{1.2}{1.29} \approx 0.93$, as $c = 1.5$ is a tighter approximation.

$\mathcal{O}(\mathcal{DB}, R_{\mathsf{min}}, \cdot)$ and $\mathcal{O}(\mathcal{DB}, R_{\mathsf{max}}, \cdot)$, respectively. Because the latter ratio is simply $\frac{R_{\mathsf{max}}}{R_{\mathsf{min}}}$, we get that Protocol 1 leaks at most a multiplicative factor of $0.77 \cdot \frac{R_{\mathsf{max}}}{R_{\mathsf{min}}}$ more compared to the ideal leakage of Functionality 1, derived in Theorem 1. ■

We empirically compare the concrete leakage of Protocol 1 to the ideal leakage in Appendix B.1 on real world data. We do so by finding values for R_{min} and R_{max} as a function of L (see Section 8.2.1 for how this is done) and applying Claim 5.

Corollary 1. The leakage of k queries to Protocol 1 is bounded by a multiplicative factor of $0.77 \cdot \frac{R_{\mathsf{max}}}{R_{\mathsf{min}}}$ more than the leakage of k queries to Functionality 1.

Proof. We only need to consider the relative leakage between k successive queries to $\mathcal{O}(\mathcal{DB}, R_{\mathsf{min}}, \cdot)$ and $\mathcal{O}(\mathcal{DB}, R_{\mathsf{max}}, \cdot)$, resulting in at most a factor of k more leakage. Another attack one might consider is to learn more information across multiple queries. While the information given by k hashes for the same vector is more specific (e.g., the set of points that have both hashes is smaller), the same effect occurs with queries to the ideal oracle. If two queries spaced by some offset return the same nearest neighbor, then that neighbor must be in the intersection of the regions of radius R_{max} centered at each query. The leakage factor captures the extra precision in our case. It also might be the case that the malicious client is able to pick queries by exploiting hashes that return zeroes. However, this is bounded by obtaining a new element of S_j (the set of all LSH digests for a vector $\boldsymbol{v_j}$; see proof of Claim 3) each time, which is captured in our concrete leakage bound. ■

B.1 Empirical leakage calculation

We chart the leakage factor in Table 7 for the real world datasets described in Table 3. For each dataset, different radii are required in order to separate near and nearest neighbors using the radix approach of Section 4.2. It is this extra precision that is captured in the leakage factor. We note that this leakage is a worst-case bound on the actual concrete leakage. Specifically, this leakage assumes that *all* points collide at the smallest radius R_{min}, when in practice the points will collide across buckets defined by radii between R_{min} and R_{max}.

Dataset	$L = 10$	$L = 30$	$L = 50$
DEEP1B	1.97×	2.79×	3.29×
MNIST	2.16×	3.23×	3.92×
GIST	3.25×	6.86×	11.32×
SIFT	3.52×	8.24×	15.7×

Table 7: Relative leakage $0.77 \cdot \frac{R_{\mathsf{max}}}{R_{\mathsf{min}}}$ versus ideal following Claim 3, evaluated on four real datasets. Leakage increases roughly proportional with the number of tables (L) given that each additional table introduces more precision in the resulting answer (assuming additional tables are used to increase accuracy). See Section 8.2.1 for more details. The 0.77 factor comes from the imprecision of the LSH computed in Lemma 1.