

Fuzzy Private Set Intersection under Density-Bounded Assumptions

Seunghun Paik
Department of Mathematics and
Research Institute for Natural Sciences
Hanyang University
whitesoonguh@hanyang.ac.kr

Jae Hong Seo✉
Department of Mathematics and
Research Institute for Natural Sciences
Hanyang University
jaehongseo@hanyang.ac.kr

***Abstract*—This paper presents a new fuzzy private set intersection (FPSI) protocol, a two-party protocol to securely identify similar items across private sets. Our design departs from the prevailing separation-based paradigm—e.g., "apart by 2δ" or the disjoint projection assumption—which either incurs exponential overhead in dimensionality or requires high dimensionality to justify the assumption. Instead, we address this tension by explicitly allowing collisions within δ-balls, but requiring that their multiplicity be bounded per axis. We formalize this new structural property as C-max collision. We theoretically analyze its plausibility on several common distributions, e.g., the uniform distribution, and real datasets, and provide an efficient algorithm to verify this assumption.**

Under the C-max collision property, we propose an efficient FPSI protocol for ℓ_∞ distance that supports both low- and high-dimensional regimes. Our construction is based on a novel abstraction that views FPSI as an oblivious intersection of membership tags, which natively handles axis-wise matching multiplicity. We capture its requirements through a new functionality called oblivious selected set intersection, along with its instantiation based on symmetric key techniques. Assuming the C-max collision property on the receiver's set, our protocol achieves computation/communication complexities scaling linearly with both parties' set sizes for a fixed C. We implement our protocol and conduct extensive evaluation. While a direct comparison is infeasible, ours achieves online efficiency comparable to state-of-the-art separation-based FPSI protocols in both low- and high-dimensional regimes.

1. Introduction

Fuzzy private set intersection (FPSI) is a two-party protocol that enables securely identifying similar items across each party's set. With the increasing needs of privacy-preserving similarity matching, e.g., contact tracing [1], ride sharing [2], deduplication [3], and biometric authentication [4], [5], various protocols have been recently proposed [6], [7], [8], [9], [10], [11], [12], [13]. More recently, several protocols have improved concrete efficiency, closing the gap between theory and practice [3], [14], [15], [16].

For efficient protocol design, existing constructions typically impose assumptions on the structure of input sets. Among them, separation-based assumptions have been the dominant design choice, such as requiring items to be at least 2δ apart [6], [7], [8], [10], [14], [15], [17] or enforcing disjoint projections across axes [6], [8], [9], [15], [16], [18]. These assumptions effectively control collisions within the δ-neighborhoods or their axis projections of set items, facilitating the design of fuzzy matching techniques such as spatial hashing [6] or fuzzy mapping [9].

Despite their impressive progress in recent years, separation-based constructions are complementary rather than overlapping. For the data dimension d and the threshold δ, protocols based on "apart by 2δ" with spatial hashing incur $\Omega(2^d)$ overhead, rendering them practical only when d is a small constant [6], [7], [8], [10], [14], [15], [17]. In contrast, while protocols from disjoint projection assumptions achieve $O(\mathrm{poly}(d))$ overhead, they are typically well justified only for high-dimensional data [8], [9], [16], [18]. This leaves a tension on the effective dimension regimes, and to our knowledge, no existing separation-based FPSI protocols can handle both regimes. This motivates us to raise the following research question:

Are separation-based assumptions inherently necessary for designing an efficient FPSI protocol?

There have been various attempts to answer this question by removing or alleviating separation-based assumptions [3], [5], [12], [13], [15]. However, existing solutions still leave the gap on the effectively supported dimension regime or require quadratic overhead $O(MN)$ in the parties' set size M, N. For example, multiple studies attempted to remove assumptions on input sets via generic two-party computation techniques [3], [5], [13]. Although recent works drastically improved their concrete efficiency, they require quadratic overhead in computation and/or communication costs. A recent work [15] showed that eliminating the separation assumption without $O(MN)$ overhead is possible, but they fall short in dealing with high-dimensional data due to their $O((\log \delta)^d)$ overhead. A concurrent work [12] attempted to relax separation-based assumptions by admitting a bounded number of overlaps between items—a type of assumption we refer to as *density-bounded*. However, they use a separation-based FPSI protocol as a black box, and thus inherit the aforementioned dimensionality issue.

1.1. Our Contribution

This paper proposes the first efficient FPSI protocol that effectively handles low- and high-dimensional regimes simultaneously. Instead of prevailing separation-based assumptions, we introduce a new density-bounded assumption that permits overlapping items per axis up to a bounded number, say C. We formalize this as the *C-max collision* property, and analyze its plausibility on various data distributions. Particularly, we show that randomly sampled items satisfy the C-max collision property with $C = O(\frac{\log d}{\log \log d})$.

Under the C-max collision property, we present a novel abstraction of the FPSI protocol based on membership tags. At a high level, for each axis, we encode the set as a set-valued function that, given a point, returns the indices (membership tags) of items whose δ-neighborhood contains that point along the axis. This reduces the fuzzy membership test in ℓ_∞ metric to an oblivious intersection of these tags across all axes. We crystallize this functionality as oblivious selected set intersection (OSSI), and instantiate it using lightweight symmetric-key cryptographic primitives. Our protocol achieves both linear communication and computation costs with respect to set sizes for a fixed C.

We implement the proposed protocol and benchmark its efficiency through experiments. In particular, as references, we also evaluate two state-of-the-art separation-based FPSI protocols tailored for low- [14] and high-dimensional [16] data. Although a strictly fair comparison is infeasible due to the incomparability of the underlying assumptions, our protocol simultaneously achieves online efficiency comparable to each reference in its respective regime. This demonstrates that separation-based assumptions are not necessary for designing an efficient FPSI protocol.

- We propose a new density-bounded assumption for efficient FPSI, called C-max collision, whose plausibility is theoretically proven, and the exact parameter C can be efficiently computed.
- We introduce a new membership tag-based abstraction of FPSI for ℓ_∞ metric, and capture its requirements via the OSSI functionality. We instantiate it from symmetric key techniques.
- We propose an efficient FPSI for ℓ_∞ metric—achieving linear overhead in parties' set sizes—under the C-max collision assumption for a fixed C, supporting both low- and high-dimensional regimes simultaneously.
- We implement and evaluate the proposed protocol, showing that it achieves comparable efficiency to state-of-the-art separation-based FPSI protocols.
- We release our source code[1] for reproducibility.

Limitation. Our FPSI construction supports the ℓ_∞ metric only, which is relevant to location-based applications such as location-based services [19]. In contrast, several prior FPSI protocols support general ℓ_p metrics [8], [12], [14], [15], [16]. Extending our construction beyond ℓ_∞ requires additional techniques, as discussed in Section 8.

1. https://doi.org/10.6084/m9.figshare.33941803

1.2. Concurrent Work

Concurrently and independently, Singh et al. [12] proposed an FPSI protocol that can handle overlapping items in δ-neighborhoods up to a bounded number, via a transformation enabling black-box use of existing separation-based FPSI protocols. While elegant in the low-dimensional regime, their construction relies on subroutines (e.g., kd-tree and underlying FPSI protocols) that become drastically inefficient as the data dimension grows. Our construction, in contrast, can effectively support high-dimensional data as well at the cost of a stronger assumption; thus, while both constructions adopt density-bounded assumptions, only ours can effectively handle both low- and high-dimensional data.

2. Technical Overview

Throughout this section, we denote $[d] := \{1, \ldots, d\}$ for $d \in \mathbb{N}$. For $u \in \mathbb{N}$, we denote the universe set $\mathcal{U}$ as the set of u-bit unsigned integers, namely, $\mathcal{U} := [0, 2^u) \cap \mathbb{Z}$.

2.1. Prior Separation-based Assumptions for FPSI

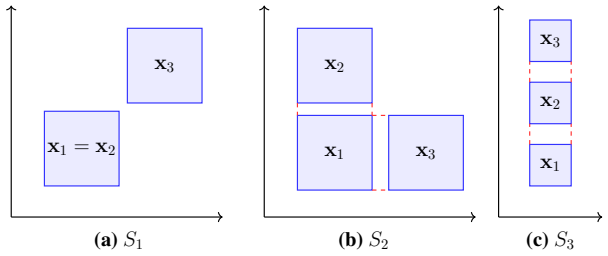
FPSI is a two-party functionality between the receiver and the sender, where the receiver privately learns the sender's items that are close enough, i.e., lie within the threshold, to some of the receiver's items. For efficient protocol design, previous studies predominantly employed two types of assumptions based on separation: "apart by 2δ" and disjoint projection. The former requires that each pair of items is far apart by 2δ so that their δ-neighborhoods do not overlap each other. Meanwhile, the latter requires that for each item, there exists at least one coordinate in which the item separates other items by at least 2δ.

Recent FPSI protocols leverage these assumptions to coarsely filter matching candidates; the effectiveness of this filtering determines their efficiency. Protocols based on "apart by 2δ" require $O(2^d)$ computation/communication costs, as their coarse filtering (e.g., spatial hashing) leaves 2^d candidates [6], [8], [14], [19]. Meanwhile, protocols based on disjoint projection enjoy $O(\mathrm{poly}(d))$ overhead by leveraging axis-wise disjoint structure, but their assumption is well-justified only for uniformly distributed, high-dimensional data [8], [9], [16], [18]. Hence, these protocols cover complementary efficiency regimes: no single separation-based protocol escapes both the $O(2^d)$ blow-up and the high-dimensionality requirement.

This study aims to overcome the tension identified above—the $O(2^d)$ overhead on one side, and the high-dimensionality requirement imposed by the assumption on the other. To this end, we depart from the prevailing separation-based paradigm and propose a new density-based design, yielding a protocol that avoids $O(2^d)$ overhead without relying on any high-dimensionality assumption.

2.2. New Density-Bounded Assumption

To avoid $O(2^d)$ overhead, we would exploit the axis-wise structure to efficiently filter out non-matches in ℓ_∞



(a) S_1 **(b)** S_2 **(c)** S_3

Figure 1: Incompatible examples between C-max collision and separation-based assumptions. (a) C-max collision does not imply "apart by 2δ". (b) C-max collision does not imply disjoint projection. (c) The separation-based assumptions do not imply C-max collision.

metric. While disjoint projection assumptions have been developed from this insight, they enforce axis-wise separability too aggressively, resulting in highly restrictive assumptions. We instead focus on axis-wise density constraints, i.e., restricting the maximum number of collisions for each axis, to leverage efficient axis-wise candidate filtering while supporting low-dimensional data seamlessly.

C-Max Collision Property. For a set $X \subset \mathcal{U}^d$, we say that X satisfies the C-max collision property of radius δ if at most C items may overlap within a δ-neighborhood for each axis. If we denote each item $x_i \in X$ as $(x_{i,1}, \ldots, x_{i,d})$, the following inequality must hold.

$$\max_{j \in [d]} \left\{ \max_{v \in \mathcal{U}} \left| \{x_i \in X \mid x_{i,j} \in [v-\delta, v+\delta] \cap \mathcal{U}\} \right| \right\} \leq C.$$

Relationship with Separation-based Assumptions. Unlike separation-based assumptions, e.g., "apart by 2δ" and disjoint projection, this property permits arbitrary proximity between items, as long as the multiplicity of overlap along each coordinate remains bounded. This makes the C-max collision property incomparable with existing separation-based assumptions, as exhibited by the following examples.

- (C-max $\not\Rightarrow$ "Apart by 2δ") Let $S_1 = \{(0,0), (0,0), (3,3)\}$ be a multiset and set $\delta = 1$, i.e., two points exactly overlap. Although there are exactly overlapping items, S_1 satisfies 2-max collision.
- (C-max $\not\Rightarrow$ disj. proj.) Let $S_2 = \{(0,0), (0,3), (3,0)\}$ and consider $\delta = 1$. Although S_2 fails to satisfy disjoint projection, it satisfies 2-max collision.

Similarly, one can easily produce the opposite case where all set items are aligned with a specific axis, e.g., $S_3 := \{(0,0), (0,3), (0,6)\}$. For $\delta = 1$, both "apart by 2δ" and disjoint projection hold owing to the 2nd axis, while failing the C-max collision property for all $C < 3$. Fig. 1 visualizes all the incompatible cases S_1, S_2, and S_3.

Plausibility Analysis based on Local Density. The plausible range of C depends on the data distribution. We analyze this by introducing the *local density* of a distribution, which measures the density of a distribution in a δ-window. For a distribution $\mathcal{D}$ over $\mathcal{U}$, we denote $p_{\mathcal{D}}(v;\delta) = \Pr_{x \sim \mathcal{D}}[x \in [v-\delta, v+\delta]]$ for $v \in \mathcal{U}$. Utilizing local density, we can obtain the desired range of C on various distributions. As in studies on disjoint projection assumptions [8], [9], [18], we show that the C-max collision property holds with high probability for randomly sampled items from $\mathcal{U}^d$. Specifically, we show that $C = O(\frac{\log d}{\log \log d})$. The full statement of the generalized theorem will be given in Section 4.2.

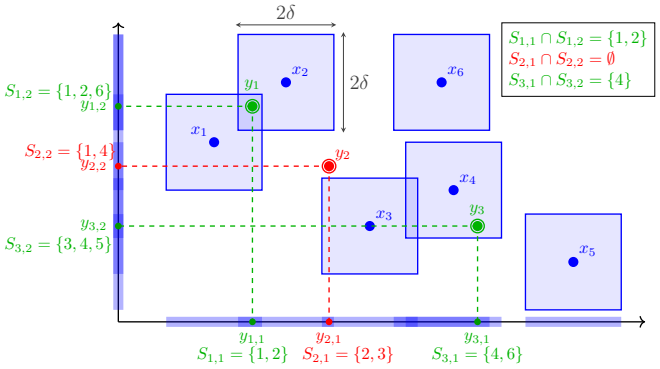


Figure 2: Visualization of our idea on $X = \{x_1, \ldots, x_6\}$ and $Y = \{y_1, y_2, y_3\}$ in $\mathcal{U}^2$. For each y_i, membership is decided by whether $S_{i,1} \cap S_{i,2}$ is empty.

Local Verification of C-Max Collision Property. In addition to our theoretical analyses, we show that the suitable C value for the given set of items can be efficiently calculated. The core idea is that if the items are sorted over $\mathcal{U}$, then a linear scan to count the number of overlapped items over M indices suffices. With radix sort, one can count the number of overlaps for each axis using this idea in time $O(udM)$. The detailed idea will be provided in Section 4.2. This suggests that the proposed assumption can be validated in advance rather than remaining purely hypothetical.

2.3. FPSI Abstraction from Membership Tags

We model the fuzzy matching in ℓ_∞ metric as the set intersection of candidate indices. We start by recalling the definition of ℓ_∞ metric: for items $x_i = (x_{i,1}, \ldots, x_{i,d})$ and $y = (y_1, \ldots, y_d)$, $\|x_i - y\|_\infty \leq \delta$ if and only if $|x_{i,j} - y_j| \leq \delta$ for all $j \in [d]$. For the set $X = \{x_i\}_{i=1}^M$, this observation motivates us to consider the index sets, say I_j, to filter out the candidates per axis: $i \in I_j$ if and only if $|x_{i,j} - y_j| \leq \delta$. With this setting, we can check $\|x_i - y\|_\infty \leq \delta$ by testing whether the corresponding i is not filtered out across all axes, i.e., $i \stackrel{?}{\in} \bigcap_{j=1}^d I_j$. Fig. 2 visualizes our idea.

We formalize the above idea by encoding X as a collection of set-valued functions $\{M_j : \mathcal{U} \to \mathcal{P}([M])\}_{j=1}^d$, called the *membership encoding* of X, such that for $v \in \mathcal{U}$,

$$M_j(v) = \{i \in [M] \mid |x_{i,j} - v| \leq \delta\}$$

We call sets in the image of M_j as *membership tags*, and denote $\mathsf{MemEnc}(X;\delta) \to \{M_j\}_{j=1}^d$ as an algorithm for computing the above membership encoding. The fuzzy

membership of y in X can be checked as follows: y is δ-close to some element in X if and only if the intersection of corresponding membership tags, i.e., $\bigcap_{j=1}^{d} M_j(y_j)$, is nonempty. If X has C-max collision property, for all $v \in \mathcal{U}$ and $j \in [d]$, $|M_j(v)| \leq C$ always holds by definition.

This idea yields the following toy *fuzzy set intersection* protocol. Suppose that the receiver $\mathcal{R}$ and the sender $\mathcal{S}$ hold sets $X, Y \subset \mathcal{U}^d$ of size M, N, respectively, and $\mathcal{R}$ wishes to learn which of its items is δ-close to $\mathcal{S}$'s items. $\mathcal{R}$ first computes $\{M_j\}_{j=1}^{d} \leftarrow \mathsf{MemEnc}(X)$ and sends $\{M_j\}_{j=1}^{d}$ to $\mathcal{S}$. For each $(y_{i,1}, \ldots, y_{i,d}) \in Y$, $i \in [N]$, $\mathcal{S}$ then computes $S_i := \bigcap_{j=1}^{d} M_j(y_{i,j})$ and returns $\{S_i\}_{i=1}^{N}$ to $\mathcal{R}$. Finally, $\mathcal{R}$ obtains the matched indices by checking whether $S_i \neq \emptyset$, and requests $\mathcal{S}$ for the corresponding y_i.

Comparison with Fuzzy Mapping Framework. Gao et al. [9] proposed the *fuzzy mapping* framework, which generalizes various FPSI protocols from separation-based assumptions [6], [8]. This framework provides a distance-aware encoding of items into sets: if two items lie within the threshold, the intersection of their encodings is nonempty. However, fuzzy mapping inherently forces a strict separation between items, as it requires that encodings of two distinct items should not overlap. Handling matching multiplicity remains open in their framework. Our approach detours this issue by encoding sets in terms of the universe, up to a bounded matching multiplicity, reflecting multiple matches via the cardinality of the intersection of membership tags.

2.4. Oblivious Selected Set Intersection (OSSI)

To make the toy protocol secure, $\mathcal{R}$ needs to hide each membership tag to $\mathcal{S}$, while $\mathcal{S}$ must not expose which points $(y_{i,j})$ are evaluated. Under these constraints, $\mathcal{R}$ should retrieve the intersection of selected membership tags; note that $M_j(y_{i,j})$ may be an empty set. We crystallize these requirements by introducing a new functionality called the oblivious selected set intersection (OSSI; $\mathcal{F}_{\mathsf{OSSI}}$). OSSI is a two-party functionality between $\mathcal{R}$ and $\mathcal{S}$; $\mathcal{R}$ inputs key-set pairs with distinct keys and $\mathcal{S}$ holds query keys. The cardinality of each set is at most some constant, say C. The functionality returns $\mathcal{R}$ to the intersection of sets corresponding to $\mathcal{S}$'s keys; if some key(s) are not matched with any $\mathcal{R}$'s key-set pairs, then $\mathcal{R}$ receives an empty set.

By incorporating $\mathcal{F}_{\mathsf{OSSI}}$ with the toy protocol, we can build a secure FPSI protocol with the oblivious transfer (OT) functionality $\mathcal{F}_{\mathsf{OT}}$. First, $\mathcal{R}$ computes the membership encoding of its set X, i.e., $\{M_j\}_{j=1}^{d} \leftarrow \mathsf{MemEnc}(X)$. For each $\mathcal{S}$'s item $y_i{=}(y_{i,1}, \ldots, y_{i,d})$, $i \in [N]$, $\mathcal{R}$ and $\mathcal{S}$ invoke $\mathcal{F}_{\mathsf{OSSI}}$ on inputs $\{(v\|j, M_j(v)) | v \in \mathcal{U}, j \in [d], M_j(v) \neq \emptyset\}$ and $(y_{i,j}\|j)_{j=1}^{d}$, respectively, returning a set S_i to $\mathcal{R}$. Here, by definition of $\mathcal{F}_{\mathsf{OSSI}}$, we can observe that $S_i = \cap_{j=1}^{d} M_j(y_{i,j})$ as in the toy protocol. After retrieving $\{S_i\}_{i=1}^{N}$, $\mathcal{R}$ sets $b_i = 1$ if $S_i \neq \emptyset$; otherwise $b_i = 0$. To retrieve $\mathcal{S}$'s item, both $\mathcal{R}$ and $\mathcal{S}$ run $\mathcal{F}_{\mathsf{OT}}$ on inputs b_i and $(\perp, y_i)$, and $\mathcal{R}$ retrieves z_i. Finally, $\{z_i \mid b_i = 1\}$ becomes the desired output. Fig. 3 visualizes the flow of the above protocol.

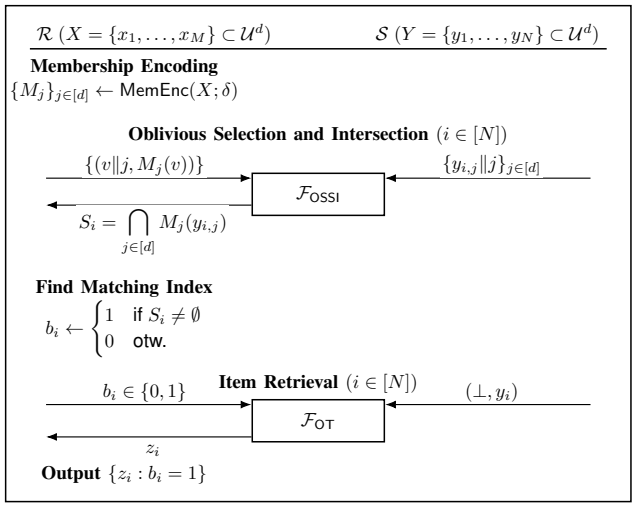


Figure 3: Overview of the proposed FPSI protocol in ℓ_∞ distance. $\delta > 0$ denotes the threshold for proximity.

2.5. Instantiating OSSI

To instantiate $\mathcal{F}_{\mathsf{OSSI}}$, we leverage an oblivious programmable pseudorandom function (OPPRF) [20] to support oblivious encoding of key-set pairs. We employ its secret-shared variant, the shared-output OPPRF (so-OPPRF) [16], which returns an additive share of a programmed value. We then adopt polynomial representations of sets to implement intersection over secret-shared membership tags.

Key-Set Pair Encoding via so-OPPRF. OPPRF is a two-party functionality, say Alice $\mathcal{A}$ and Bob $\mathcal{B}$, which allows $\mathcal{A}$ to program chosen input-output pairs of a PRF and $\mathcal{B}$ to query input points for a PRF evaluation. In its shared-output variant of OPPRF [16] ($\mathcal{F}_{\mathsf{soOPPRF}}$), both parties obtain additive shares of the PRF evaluation result from $\mathcal{B}$'s queries. Such shares can be utilized for downstream secure two-party computation on programmed values. Yang et al. [16] showed that so-OPPRF can be constructed from an oblivious key-value store (OKVS) [21] and a shared-output OPRF [22].

We handle the key-set pairs in $\mathcal{R}$'s OSSI input as key-item pairs by enumerating the set items. For each key-set pair (k_i, S_i), let us denote $m_i := |S_i|$ and $s_{i,1}, \ldots, s_{i,m_i}$ as an enumeration of S_i. Then $\mathcal{R}$ encodes (k_i, S_i) as m_i key-item tuples, namely, $\{(k_i\|l, s_{i,l})\}_{l=1}^{m_i}$, and serves as $\mathcal{A}$ in so-OPPRF. The OSSI $\mathcal{S}$, serving as $\mathcal{B}$ in so-OPPRF, could then retrieve the programmed S_i by querying $k_i\|1, \ldots, k_i\|m_i$. However, $\mathcal{S}$ does not know the exact m_i; so it queries all C possible keys $k_i\|1, \ldots, k_i\|C$ and receives additional $C - m_i$ dummy values. Furthermore, it cannot distinguish which values are dummy, as the received values are secret-shared.

While the above encoding may be counterintuitive—$\mathcal{R}$ and $\mathcal{S}$ learn shares of *noisy* sets with dummy elements—such dummies in fact do not affect the result with high probability. We prove that, for the output space $\mathbb{F}$ of so-OPPRF, the probability that dummy values change the intersection result is at most $\frac{C^2}{|\mathbb{F}|} + \frac{C^{d+1}}{|\mathbb{F}|^d}$ (Theorem 3). Thus, we can safely proceed with noisy sets when $|\mathbb{F}|$ is large enough.

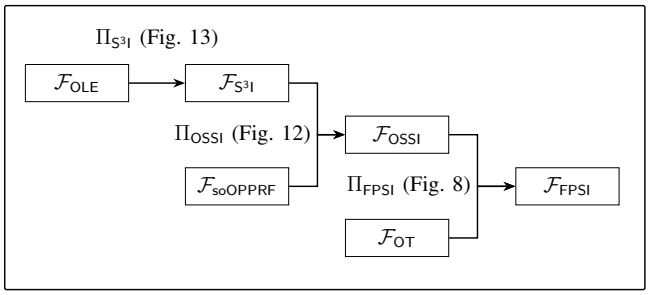


Figure 4: Modular view of our FPSI construction. $\Pi_{\mathsf{S^3I}}$, Π_{OSSI}, Π_{FPSI} are protocols that connect each functionality.

Secret-Shared Set Intersection ($\mathsf{S^3I}$). After invoking so-OPPRF, $\mathcal{R}$ and $\mathcal{S}$ obtain additive shares of elements in selected sets with dummy elements, and they need to compute the intersection of these noisy sets. Essentially, we need a functionality to compute the intersections of secret-shared sets, which we call secret-shared set intersection ($\mathsf{S^3I}$; $\mathcal{F}_{\mathsf{S^3I}}$).

We implement $\mathcal{F}_{\mathsf{S^3I}}$ by representing each set as a polynomial. Let S_j be the selected set by $\mathcal{S}$ in j^{th} index. We encode S_j as $f_{S_j}(X) := \prod_{z\in S_j}(X-z) \in \mathbb{F}[X]$. It is known that the random linear combination $\sum_{j=1}^{d}\gamma_j(X)f_{S_j}(X)$ for $\gamma_j(X) \in \mathbb{F}[X]^{\le C-1}$ encodes $S_I := \cap_{j=1}^{d} S_j$ multiplied by some random polynomial in $\mathbb{F}[X]^{\le 2C-|S_I|-1}$ [23]. One can show that such a random linear combination can be computed in a secret-shared form with $O(dC\log C)$ oblivious linear evaluations (OLE; $\mathcal{F}_{\mathsf{OLE}}$) over $\mathbb{F}$ (Section 6.3).

With this encoding, we employ a gcd trick by Yeo and Ying [24], which finds the intersection by computing the gcd of two intersection polynomials from independent randomness. Precisely, for random polynomials $\{\gamma_{j,1}(X)\}_{j=1}^{d}$ and $\{\gamma_{j,2}(X)\}_{j=1}^{d}$ sampled from $\mathbb{F}[X]^{\le C-1}$, the following equality holds with failure probability at most $\frac{1}{|\mathbb{F}|-1}$:

$$\gcd\Big(\sum_{j=1}^{d}\gamma_{j,1}(X)f_{S_j}(X), \sum_{j=1}^{d}\gamma_{j,2}(X)f_{S_j}(X)\Big) = \prod_{z\in S_I}(X-z)$$

Putting Them All Together. Using $\mathcal{F}_{\mathsf{soOPPRF}}$ and $\mathcal{F}_{\mathsf{S^3I}}$ with our instantiation, we can implement $\mathcal{F}_{\mathsf{OSSI}}$ by applying the above-mentioned key-set pairs encoding with so-OPPRF and $\mathsf{S^3I}$ with noisy sets. Therefore, from the observation in Section 2.4, we can securely implement the FPSI functionality $\mathcal{F}_{\mathsf{FPSI}}$ in $(\mathcal{F}_{\mathsf{soOPPRF}}, \mathcal{F}_{\mathsf{OLE}}, \mathcal{F}_{\mathsf{OT}})$-hybrid model. We summarize the dependency between functionalities in Fig. 4, with reference sub-protocols in our instantiation.

3. Preliminaries

Notation. We denote $\|\cdot\|_\infty$ as the standard ℓ_∞ metric, i.e., $\|(x_i)_{i=1}^{d}\|_\infty := \max_{i\in[d]}\{|x_i|\}$. log denotes the logarithm of base 2. For a set S, we denote $\mathcal{P}(S)$ as the collection of all subsets of S. For a finite set S, we denote $x \xleftarrow{\$} S$ as the uniform sampling from S. For a field $\mathbb{F}$ and $k \in \mathbb{N}$, we denote $\mathbb{F}[X]$ as the set of polynomials, and $\mathbb{F}[X]^{\le k}$ as its subset whose degree is at most k. We denote $\cup$ as the usual set union, whereas $\uplus$ denotes the multiset union. We treat $\cap$ as the multiset intersection, i.e., min-multiplicity. For a multiset S, we denote $z \in S$ as multiset enumeration, i.e., $S = \{z \mid z \in S\}$; if we need to view supports only, we will use $z \in \mathrm{Supp}(S)$ instead.

Parameters: Two parties, the receiver $\mathcal{R}$ and the sender $\mathcal{S}$. The universe set $\mathcal{U}$ and the data dimension $d \in \mathbb{N}$. The threshold $\delta > 0$. $\mathcal{R}$ and $\mathcal{S}$'s set sizes M, N, respectively.
Inputs:
- $\mathcal{R}$: A set $X = \{x_1, \ldots, x_M\} \subset \mathcal{U}^d$.
- $\mathcal{S}$: A set $Y = \{y_1, \ldots, y_N\} \subset \mathcal{U}^d$.

Outputs:
- $\mathcal{R}$: A set $\{y_i \in Y \mid \exists x_j \in X \text{ s.t. } \|x_j - y_i\|_\infty \le \delta\}$.
- $\mathcal{S}$: Nothing.

Figure 5: The ideal functionality $\mathcal{F}_{\mathsf{FPSI}}$ for ℓ_∞ metric.

Fuzzy Private Set Intersection. We consider the receiver $(\mathcal{R})$ and the sender $(\mathcal{S})$ scenario where the $\mathcal{R}$ only retrieves the protocol output. In FPSI, $\mathcal{R}$ and $\mathcal{S}$ hold sets X, Y, respectively, consisting of d-dimensional items in $\mathcal{U}^d$. For a distance metric d in $\mathcal{U}^d$ and a matching threshold δ, $\mathcal{R}$ wishes to retrieve $\mathcal{S}$'s item that is δ-close to at least one of $\mathcal{R}$'s items in d, i.e., $\{y \in Y \mid \exists x \in X \text{ s.t. } \mathsf{d}(x,y) \le \delta\}$. Among various choices of d, we focus on the ℓ_∞ metric, which measures the maximum distance between each coordinate. Fig. 5 describes the ideal functionality $\mathcal{F}_{\mathsf{FPSI}}$.

Semi-Honest Adversary Model. For a two-party functionality $\mathcal{F}$ on parties P_1, P_2, a protocol Π securely realizes $\mathcal{F}$ under the semi-honest model if there is a simulator Sim that produces the view (View) of each party in Π from the inputs/outputs of $\mathcal{F}$ [25]. Here, View includes protocol input, internal randomness, and the transcript. Formally, for $i \in \{1,2\}$ the following two distributions must be computationally indistinguishable:

$$\{\mathsf{Sim}(i, x_i, f_i(x_1,x_2), 1^\lambda)\} \approx_c \{\mathsf{View}(\Pi, P_i, x_1, x_2, 1^\lambda)\}$$

Here, f_i denote P_i's ideal functionality output.

Useful Lemmas. We will use the following lemmas:

Lemma 1 (Chernoff Bound for Binomial Distribution [26])**.** *Let $\mathfrak{X}$ be a random variable that follows a binomial distribution* $\mathrm{Bin}(n,p)$. *We denote $\mu = np$. Then for $k = (1+\xi)\mu$ with $\xi > 0$, $\Pr[\mathfrak{X} > k] < \left(\frac{e^\xi}{(1+\xi)^{1+\xi}}\right)^\mu$. In particular, if $k > e\mu$, this bound can be simplified by $\Pr[\mathfrak{X} > k] < \left(\frac{e\mu}{k}\right)^k$.*

4. New Density-Bounded Assumption

We introduce a novel density-bounded assumption called the C-max collision, with its plausibility analysis. Later, we will design our FPSI protocol under this assumption.

4.1. C-Max Collision Property

As briefly discussed in Section 2.2, we depart from the prevailing separation-based paradigms to axis-wise density, aiming to mitigate the tension on the dimension d. Accordingly, we introduce the C-max collision property that restricts the maximum number of collisions of set items on an axis to at most C for all axes, as formally stated below:

Definition 1. *Let $\mathcal{U} := [0, 2^u) \cap \mathbb{Z}$ be the universe set. For parameters $C, \delta \in \mathbb{N}$ and a set $X = \{x_1, \ldots, x_M\} \subset \mathcal{U}^d$, we say that X has a C-max collision property of radius δ if the following inequality holds:*

$$\max_{j\in[d]} \left\{ \max_{v\in\mathcal{U}} |\{x_i \in X : x_{i,j} \in [v-\delta, v+\delta] \cap \mathcal{U}\}| \right\} \leq C,$$

where $x_{i,j}$ denotes the jth coordinate of $x_i \in \mathcal{U}^d$.

Note that this property trivially holds when $C \geq M$. In addition, when $C = 1$, this property reduces to the global disjoint projection assumption [6], which has been described as rather artificial in prior works [6], [8].

Relationship with Separation-Based Assumptions. Unlike separation-based assumptions, the new assumption enforces bounded density, i.e., overlaps on δ-neighborhoods may happen. This discrepancy makes it incomparable with separation-based ones, as we visualized in Fig. 1. We remark that we can view the global disjoint projection assumption [6] as a 1-max collision assumption, and our assumption always holds whenever $C \geq M$. From this perspective, we can interpret the C-max collision assumption as a middle-ground between the global disjoint projection and "assumption-free" settings in terms of axis-wise density.

4.2. Plausibility Analysis of the Assumption

We now analyze the plausibility of the C-max collision property, deriving a suitable C when each item in a set X follows some specific distribution. In this section, we focus on theoretical analysis. Empirical analyses on concrete parameters will be provided in Section 7.1. We defer full proofs of our theoretical results to Appendix B.1– B.3; we outline the key proof ideas below. We view the C-max collision property as counting the number of items in a δ-window. Hence, we introduce *local density* to capture the local concentration of the given distribution.

Definition 2. *Let $\mathcal{D}$ be a distribution over $\mathcal{U}$. For $v \in \mathcal{U}$ and $\delta > 0$, we define the local density of $\mathcal{D}$ at v of radius δ as $p_{\mathcal{D}}(v; \delta) := \Pr_{x\sim\mathcal{D}}[x \in [v-\delta, v+\delta] \cap \mathcal{U}]$.*

From this, we model the number of items in the given δ-window through a random variable following a binomial distribution. With the Chernoff bound in Lemma 1, we can derive the probability that the number of collided items exceeds C, i.e., the C-max collision property fails. From this idea, we prove the following statement:

Theorem 1. *Let $\mathcal{U}$ be the universe and $d \in \mathbb{N}$ be the data dimension. For a distribution $\mathcal{D}$ over $\mathcal{U}^d$, we denote $\mathcal{D}_j$ as its marginal distribution on j^{th} coordinate. For $x_1, \ldots, x_M$ i.i.d., sampled from $\mathcal{D}$ and $C \in \mathbb{N}$, the following inequality holds for the set $X = \{x_1, \ldots x_M\}$ and $C \geq eM \max_{j\in[d], v\in\mathcal{U}} p_{\mathcal{D}_j}(v; \delta)$:*

$$\Pr[X \text{ is } C\text{-max collision}] \geq 1 - \sum_{j=1}^{d} \sum_{v\in\mathcal{U}} \left(\frac{eMp_{\mathcal{D}_j}(v; \delta)}{C} \right)^C$$

With Theorem 1, we can compute lower bounds of C for various distributions from their local densities. For example, for the uniform distribution $\mathcal{D}_{\text{Unif}}$ over $\mathcal{U}$, we observe that $p_{\mathcal{D}_{\text{Unif}}}(v; \delta) = \frac{2\delta+1}{2^u}$ for any $v \in \mathcal{U}$. Plugging it to Theorem 1 gives the following corollary, which guarantees that the C-max collision assumption holds for uniformly distributed data when $C = O(\frac{\log d}{\log\log d})$. We provide generalized bounds of C for arbitrary distributions in Appendix C.

Corollary 1. *For $A = \kappa + u + \log d$, $\mu = \frac{M(2\delta+1)}{2^u}$ such that $A > \frac{e^2\mu}{\ln 2}$, and the Lambert W function, it suffices to choose $C \geq \max\left\{\mu e, \frac{A \ln 2}{W(A\mu^{-1}e^{-1}\ln 2)}\right\}$ to satisfy C-max collision assumption with a probability at least $1 - 2^{-\kappa}$ for a set that consists of M elements randomly sampled from $\mathcal{U}^d$.*

Local Verifiability of C-Max Collision Property. While we theoretically analyzed the plausibility of the C-max collision property, the suitable C value for the given data can also be efficiently calculated. More precisely, for an axis j, suppose that the jth components of set items are already sorted, i.e., $x_{1,j} \leq x_{2,j} \leq \cdots \leq x_{M,j}$. Here, we can count the number of items in a 2δ-interval by computing the maximum difference in indices of items in the interval. Hence, finding the maximum difference in two indices whose corresponding items are 2δ-close is sufficient, as stated by the following lemma:

Lemma 2. *For sorted $z_1 \leq \cdots \leq z_M \in \mathcal{U}$ and any $\delta > 0$,*

$$\max_{v\in\mathcal{U}} \left| \{\, i \in [M] : z_i \in [v-\delta,\, v+\delta] \,\} \right| = \max_{i\geq j,\ z_i - z_j \leq 2\delta} (i-j+1).$$

Thus, a linear scan over M sorted items suffices; radix sort enables such sorting with $O(udM)$ comparisons in total. The detailed algorithm is given in Appendix D.

5. FPSI with Membership Tags

5.1. Membership Encodings of Sets

For a set $X = \{x_1, \ldots, x_M\} \subset \mathcal{U}^d$, let us denote $x_i = (x_{i,j})_{j=1}^d \in \mathcal{U}^d$ for $i \in [M]$. For each axis $j \in [d]$, consider a set-valued function $M_j : \mathcal{U} \to \mathcal{P}([M])$ defined as

$$M_j(v) = \{i \in [M] : v \in [x_{i,j} - \delta, x_{i,j} + \delta] \cap \mathcal{U}\},$$

where $\mathcal{P}([M])$ denotes the collection of all subsets of $[M]$. We call the collection $\{M_j\}_{j=1}^d$ the *membership encoding* of the set X of radius δ. From the definition, we can straightforwardly deduce the following lemma:

Parameters: Two parties, the receiver $\mathcal{R}$ and the sender $\mathcal{S}$. The key space $\mathcal{K}$ and a field $\mathbb{F}$. The upper bound C of each $\mathcal{R}$'s set. The total cardinality V of $\mathcal{R}$'s sets and the number d of query keys by $\mathcal{S}$.
Inputs:

- $\mathcal{R}$: Q key-set pairs $\{(k_i, S_i)\}_{i=1}^{Q} \subset \mathcal{K} \times \mathcal{P}(\mathbb{F})$ with distinct keys $k_i \in \mathcal{K}$, $|S_i| \leq C$, and $\sum_{i=1}^{Q} |S_i| = V$.
- $\mathcal{S}$: d keys $\hat{k}_1, \ldots, \hat{k}_d$.

Outputs: For $i \in [d]$, define

$$\hat{S}_i = \begin{cases} S_j & \text{if } \hat{k}_i = k_j \text{ for some } j \in [Q], \\ \emptyset & \text{otherwise.} \end{cases}$$

- $\mathcal{R}$: The intersection $\bigcap_{i=1}^{d} \hat{S}_i$; $\mathcal{S}$: Nothing.

Figure 6: The ideal functionality $\mathcal{F}_{\mathsf{OSSI}}$.

Lemma 3. *Let $X = \{x_1, \ldots, x_M\} \subset \mathcal{U}^d$ be a set and $\{M_j\}_{j=1}^{d} \leftarrow \mathsf{MemEnc}(X; \delta)$ for the threshold $\delta > 0$. For $y_k = (y_{k,1}, \ldots, y_{k,d}) \in \mathcal{U}^d$ and $i \in [M]$, we have that $\|x_i - y_k\|_\infty \leq \delta \iff i \in \bigcap_{j=1}^{d} M_j(y_{k,j})$.*

Proof. $\|x_i - y_k\|_\infty \leq \delta \iff \forall j \in [d], |x_{i,j} - y_{k,j}| \leq \delta$, i.e., $y_{k,j} \in [x_{i,j} - \delta, x_{i,j} + \delta] \cap \mathcal{U}$ and $i \in M_j(y_{k,j})$. □

Hence, the fuzzy matching of y_k over the set X reduces to testing the non-emptiness of an intersection of d coordinate-wise membership encodings. In addition, if we assume the C-max collision property on the set X, each membership tag has a size at most C.

5.2. Oblivious Selected Set Intersection

As briefly discussed in Section 2.3, our membership tag-based abstraction yields the following fuzzy set intersection protocol: (i) $\mathcal{R}$ computes the membership encodings $\{M_j\}$ and sends them to $\mathcal{S}$, (ii) $\mathcal{S}$ finds membership tags corresponding to its input and returns their intersections to $\mathcal{R}$, and (iii) $\mathcal{R}$ identifies which item indices of $\mathcal{S}$ are δ-close to $\mathcal{R}$'s items. Finally, $\mathcal{R}$ can request the matched items from $\mathcal{S}$ using the obtained indices.

To make this protocol secure, we introduce a novel functionality called oblivious selected set intersection (OSSI). The key insight is to view the evaluation of the membership encodings as a lookup in a key-set table that maps each evaluation point to its corresponding membership tag. Hence, we can conceptualize the above procedure as the functionality where $\mathcal{R}$ provides key-set pairs and $\mathcal{S}$ provides some keys, returning $\mathcal{R}$ to the intersection of looked-up sets. If some $\mathcal{S}$'s key does not match any of $\mathcal{R}$'s keys, then $\mathcal{R}$ obtains an empty set. During this process, $\mathcal{R}$ must not learn which key-value pairs are being looked up, and $\mathcal{S}$ must not be able to see the naïve membership tags, as well as their intersections. Furthermore, to prevent $\mathcal{S}$ from inferring which $\mathcal{R}$'s keys are encoded, it must be impossible for $\mathcal{S}$ to identify which keys are valid. Fig. 6 describes the ideal functionality $\mathcal{F}_{\mathsf{OSSI}}$ that incorporates all the requirements discussed above.

Parameters: Two parties, the receiver $\mathcal{R}$ and the sender $\mathcal{S}$. A space $\mathcal{Y}$ where $\mathcal{S}$'s item lies in.
Inputs:

- $\mathcal{R}$: A bit $b \in \{0, 1\}$.
- $\mathcal{S}$: Two items $y_0, y_1 \in \mathcal{Y}$.

Outputs:

- $\mathcal{R}$: An item $y_b \in \mathcal{Y}$; $\mathcal{S}$: Nothing.

Figure 7: The ideal functionality $\mathcal{F}_{\mathsf{OT}}$.

Parameters: Two parties, the receiver $\mathcal{R}$ and the sender $\mathcal{S}$. The universe set $\mathcal{U}$ and the data dimension $d \in \mathbb{N}$. The threshold $\delta > 0$. $\mathcal{R}$ and $\mathcal{S}$'s set sizes M, N, respectively.
Inputs:

- $\mathcal{R}$: A set $X = \{x_1, \ldots, x_M\} \subset \mathcal{U}^d$.
- $\mathcal{S}$: A set $Y = \{y_1, \ldots, y_N\} \subset \mathcal{U}^d$.

Procedure:

(1) $\mathcal{R}$ parses $x_i = (x_{i,1}, \ldots, x_{i,d})$ and computes $\{M_j\}_{j=1}^{d} \leftarrow \mathsf{MemEnc}(X; \delta)$.
(2) $\mathcal{S}$ parses $y_i = (y_{i,j})_{j=1}^{d}$ and samples a random permutation $\pi : [N] \to [N]$. Then for each $i \in [N]$, $\mathcal{R}$ and $\mathcal{S}$ invoke $\mathcal{F}_{\mathsf{OSSI}}$ on inputs

$$\{(v \| j, M_j(v)) \mid j \in [d], M_j(v) \neq \emptyset\},$$

and $\{y_{\pi(i),j} \| j\}_{j=1}^{d}$, respectively. Finally, $\mathcal{R}$ receives sets $\{S_i\}_{i=1}^{N}$.
(3) $\mathcal{R}$ sets a bit $b_i = 1$ if $S_i \neq \emptyset$; $b_i = 0$ otherwise.
(4) For each $i \in [N]$, $\mathcal{R}$ and $\mathcal{S}$ invoke $\mathcal{F}_{\mathsf{OT}}$ on inputs b_i and $(\perp, y_{\pi(i)})$, respectively. $\mathcal{R}$ then obtains z_i.
(5) $\mathcal{R}$ outputs $\{z_i \mid b_i = 1\}$.

Figure 8: Protocol Π_{FPSI} in $(\mathcal{F}_{\mathsf{OSSI}}, \mathcal{F}_{\mathsf{OT}})$-hybrid model.

5.3. FPSI from OSSI with C-Max Collision

With $\mathcal{F}_{\mathsf{OSSI}}$, we can securely implement the toy protocol by setting its key-set pairs (k_i, S_i) of $\mathcal{R}$ as $(v \| j, M_j(v))$ such that $M_j(v) \neq \emptyset$ for $v \in \mathcal{U}$, $j \in [d]$. The total cardinality of the sets is $\sum_{v \in \mathcal{U}} \sum_{j=1}^{d} |M_j(v)| = (2\delta + 1)dM$, as each item contributes exactly $(2\delta+1)$ items per axis to the membership encoding[2]. For each $\mathcal{S}$'s item $y_i = (y_{i,1}, \ldots, y_{i,d})$, $\mathcal{R}$ and $\mathcal{S}$ invoke $\mathcal{F}_{\mathsf{OSSI}}$ with the $\mathcal{R}$'s prepared key-set pairs and $\mathcal{S}$'s keys $\{y_{i,j} \| j\}_{j=1}^{d}$. After invoking OSSI with membership tags, $\mathcal{R}$ learns $\bigcap_{j=1}^{d} M_j(y_{i,j})$. By Lemma 3, this intersection contains the indices whose corresponding $\mathcal{S}$'s items are δ-close to some of $\mathcal{R}$'s items. Finally, to retrieve such $\mathcal{S}$'s items, $\mathcal{R}$ can utilize 1-out-of-2 oblivious transfer (OT) defined over $\mathcal{U}^d$, whose functionality is described in Fig. 7. Thus, we can implement $\mathcal{F}_{\mathsf{FPSI}}$ via the protocol Π_{FPSI} in Fig. 8, whose security proof is deferred to Appendix B.4.

Theorem 2. *The protocol Π_{FPSI} in Fig. 8 securely implements $\mathcal{F}_{\mathsf{FPSI}}$ against semi-honest adversaries in $(\mathcal{F}_{\mathsf{OSSI}}, \mathcal{F}_{\mathsf{OT}})$-hybrid model.*

2. For simplicity, we ignore the case where the δ-interval is truncated at the $\mathcal{U}$'s boundary; this can be addressed by padding dummy key-set pairs.

6. Our OSSI Construction

We now present our implementation of $\mathcal{F}_{\mathsf{OSSI}}$. We defer the full proofs of our core theorems to Appendix B.5–B.8.

6.1. Cryptographic Building Blocks

Shared-Output OPPRF. We use the shared-output OPPRF (so-OPPRF) [16] described in Section 2.5, whose ideal functionality $\mathcal{F}_{\mathsf{soOPPRF}}$ is given in Fig. 9. Yang et al. showed that so-OPPRF can be built upon an oblivious key-value store (OKVS) [21] and so-OPRF [22].

(Ring) Oblivious Linear Evaluation. OLE ($\mathcal{F}_{\mathsf{OLE}}$) [27] enables obtaining an additive share of the product $a \cdot b$ of two field elements. Its ring variant, called rOLE ($\mathcal{F}_{\mathsf{rOLE}}$) [28], instead lets $\mathcal{R}$ and $\mathcal{S}$ hold $f_1, f_2 \in \mathbb{F}[X]$, respectively, s.t. $\deg f_1 \leq m$ and $\deg f_2 \leq n$, and provides an additive share of $f_1(X)f_2(X)$. When $f_1(X)$ and $f_2(X)$ are secret-shared, we call it secret-shared rOLE (SSrOLE). Fig. 10 provides its ideal functionality $\mathcal{F}^{m,n}_{\mathsf{SSrOLE}}$. SSrOLE can be constructed from OLEs, i.e., invoking 2 OLEs at $m+n+1$ evaluation points. We provide the details in Appendix A.

6.2. Instantiating OSSI from so-OPPRF and $\mathsf{S^3I}$

Encoding Oblivious Key-Set Pairs. For key-set pairs $\{(k_i, S_i)\}_{i=1}^{Q}$, $\mathcal{R}$ first enumerates the elements of S_i as $s_{i,1}, \ldots, s_{i,m_i}$, where $m_i := |S_i| \leq C$. Then $\mathcal{R}$ builds key-value tuples $\{(k_i \| l, s_{i,l})\}$, $i \in [Q]$ and $l \in [m_i]$. Note that the number of tuples is exactly $\sum_{i=1}^{Q} |S_i|$. On the other hand, $\mathcal{S}$ expands its input keys $\hat{k}_1, \ldots, \hat{k}_d$ to $\hat{k}_j \| l$ for $j \in [d]$ and $l \in [C]$, i.e., Cd keys in total. $\mathcal{R}$ and $\mathcal{S}$ then invoke $\mathcal{F}_{\mathsf{soOPPRF}}$, obtaining $w^{\mathcal{R}}_{j,l}$ and $w^{\mathcal{S}}_{j,l}$ for $j \in [d]$ and $l \in [C]$. If $\hat{k}_j$ was programmed by $\mathcal{R}$, say $\hat{k}_j = k_i$ for some $i \in [Q]$, then $\mathcal{R}$ and $\mathcal{S}$ retrieve additive shares of $s_{i,l}$, i.e.,

$$w^{\mathcal{R}}_{j,l} + w^{\mathcal{S}}_{j,l} = \begin{cases} s_{i,l} & \text{if } \exists i \in [Q] \text{ s.t. } \hat{k}_j = k_i \wedge l \leq m_i \\ r \xleftarrow{\$} \mathbb{F} & \text{otherwise.} \end{cases}$$

From the above approach, $\mathcal{S}$ must expand each $\hat{k}_j$ to C points regardless of the cardinality of the actual set S_i. This yields $C - |S_i|$ dummy values even when $\hat{k}_j = k_i$ for some $i \in [Q]$. We analyze their effect by showing that they do not affect the intersection result with high probability.

Theorem 3. *Let $\mathbb{F}$ be a field and $C \in \mathbb{N}$. For $d \geq 2$ sets $\{A_i\}_{i=1}^{d}$ such that $|A_i| \leq C$, sample multisets $\{B_i\}_{i=1}^{d}$ s.t.*

$$B_i \leftarrow \begin{cases} \{b_j \xleftarrow{\$} \mathbb{F} \mid j = 1, \ldots, C - |A_i|\} & \text{if } |A_i| < C, \\ \emptyset & \text{otherwise} \end{cases}$$

Then $\Pr[\bigcap_{i=1}^{d} A_i \neq \bigcap_{i=1}^{d} (A_i \uplus B_i)] \leq \frac{C^2}{|\mathbb{F}|} + \frac{C^{d+1}}{|\mathbb{F}|^d}$, where the probability runs over the choice of $B_1, \ldots, B_d$.

Here, A_i denotes the set of programmed items, while B_i denotes the (multi-)set of dummies. We prove this theorem by analyzing two failure modes where dummy elements expand $\cap_{i=1}^{d} A_i$, or increase the intersection multiplicity.

Parameters: Two parties, Alice $\mathcal{A}$ and Bob $\mathcal{B}$. The key space $\mathcal{K}$ and a field $\mathbb{F}$. The number of key-value pairs V of $\mathcal{A}$ and query keys T of $\mathcal{B}$.
Inputs:
- $\mathcal{A}$: A set of key-value pairs $\{(k_i, v_i)\}_{i=1}^{V}$ with distinct keys $k_i \in \mathcal{K}$ and values $v_i \in \mathbb{F}$.
- $\mathcal{B}$: A tuple of query keys $(\hat{k}_1, \ldots, \hat{k}_T)$ with $\hat{k}_j \in \mathcal{K}$.

Outputs: Sample a random function $F : \mathcal{K} \to \mathbb{F}$ subject to $F(k_i) = v_i$, $\forall i \in [V]$. For each $j \in [T]$, $w_j \leftarrow F(\hat{k}_j)$, $w^{\mathcal{A}}_j \xleftarrow{\$} \mathbb{F}$, and $w^{\mathcal{B}}_j \leftarrow w_j - w^{\mathcal{A}}_j$.
- $\mathcal{A}$: $(w^{\mathcal{A}}_1, \ldots, w^{\mathcal{A}}_T)$.
- $\mathcal{B}$: $(w^{\mathcal{B}}_1, \ldots, w^{\mathcal{B}}_T)$.

Figure 9: The ideal functionality $\mathcal{F}_{\mathsf{soOPPRF}}$.

Parameters: Two parties, the receiver $\mathcal{R}$ and the sender $\mathcal{S}$. A field $\mathbb{F}$ and polynomial degrees m, n.
Inputs:
- $\mathcal{R}$: $f^{\mathcal{R}}_1(X) \in \mathbb{F}[X]^{\leq m}$ and $f^{\mathcal{R}}_2(X) \in \mathbb{F}[X]^{\leq n}$.
- $\mathcal{S}$: $f^{\mathcal{S}}_1(X) \in \mathbb{F}[X]^{\leq m}$ and $f^{\mathcal{S}}_2(X) \in \mathbb{F}[X]^{\leq n}$.

Outputs: $f(X) \leftarrow (f^{\mathcal{R}}_1(X) + f^{\mathcal{S}}_1(X))(f^{\mathcal{R}}_2(X) + f^{\mathcal{S}}_2(X))$ and make additive shares $f^{\mathcal{R}}(X) \xleftarrow{\$} \mathbb{F}[X]^{\leq m+n}$, $f^{\mathcal{S}}(X) \leftarrow f(X) - f^{\mathcal{R}}(X)$.
- $\mathcal{R}$: The polynomial $f^{\mathcal{R}}(X)$.
- $\mathcal{S}$: The polynomial $f^{\mathcal{S}}(X)$.

Figure 10: The ideal functionality $\mathcal{F}^{m,n}_{\mathsf{SSrOLE}}$.

Secret-Shared Set Intersection. Formally, for a (multi-)set $W = \{w_1, w_2, \ldots, w_C\}$, let us define the additive share of W as a tuple of ordered sets (W^+, W^-):

$$W^+ = (w^+_1, \ldots, w^+_C),\ W^- = (w^-_1, \ldots, w^-_C)$$
$$\forall i \in [C],\ w^+_i + w^-_i = w_i.$$

With this notation, we can regard the so-OPPRF outputs $(w^{\mathcal{R}}_{j,1}, \ldots, w^{\mathcal{R}}_{j,C})$ and $(w^{\mathcal{S}}_{j,1}, \ldots, w^{\mathcal{S}}_{j,C})$ for $\mathcal{R}$ and $\mathcal{S}$, respectively, as the additive share of S_i with dummy values.

Recall that $\mathcal{F}_{\mathsf{OSSI}}$ should output the intersection of shared sets to $\mathcal{R}$. We thus employ a functionality to compute the set intersection in a shared form. For sets $W_1, \ldots, W_d$ such that $|W_j| = C$ and their shares $(W^{\mathcal{R}}_j, W^{\mathcal{S}}_j)$, $j \in [d]$, assume that $\mathcal{R}$ holds $W^{\mathcal{R}}_j$ and $\mathcal{S}$ holds $W^{\mathcal{S}}_j$. The functionality outputs $\mathcal{R}$ to $\bigcap_{j=1}^{d} W_j$ from inputs $(W^{\mathcal{R}}_j)_{j=1}^{d}$ and $(W^{\mathcal{S}}_j)_{j=1}^{d}$. We call this functionality a secret-shared set intersection ($\mathsf{S^3I}$), whose ideal functionality $\mathcal{F}_{\mathsf{S^3I}}$ is described in Fig. 11.

Instantiating OSSI. From these ideas, we can securely implement $\mathcal{F}_{\mathsf{OSSI}}$ via the protocol Π_{OSSI} described in Fig. 12. Thanks to Theorem 3, invoking $\mathcal{F}_{\mathsf{S^3I}}$ with secret-shared sets in Π_{OSSI} yields the same output set as $\mathcal{F}_{\mathsf{OSSI}}$, except with probability at most $\frac{C^2}{|\mathbb{F}|} + \frac{C^{d+1}}{|\mathbb{F}|^d}$.

Theorem 4. *The protocol Π_{OSSI} in Fig. 12 securely implements $\mathcal{F}_{\mathsf{OSSI}}$ against semi-honest adversaries in $(\mathcal{F}_{\mathsf{soOPPRF}}, \mathcal{F}_{\mathsf{S^3I}})$-hybrid model.*

Parameters: Two parties, the receiver $\mathcal{R}$ and the sender $\mathcal{S}$, and a field $\mathbb{F}$. The number of sets d, and their cardinality $C \in \mathbb{N}$.
Inputs:

- $\mathcal{R}$: d ordered sets $W_1^{\mathcal{R}}, \ldots, W_d^{\mathcal{R}}$ of cardinality C.
- $\mathcal{S}$: d ordered sets $W_1^{\mathcal{S}}, \ldots, W_d^{\mathcal{S}}$ of cardinality C.

Outputs: Parse $W_j^{\mathcal{R}} \to (w_{j,l}^{\mathcal{R}})_{l=1}^{C}$ and $W_j^{\mathcal{S}} \to (w_{j,l}^{\mathcal{S}})_{l=1}^{C}$, and set $W_j \leftarrow \{w_{j,l}^{\mathcal{R}} + w_{j,l}^{\mathcal{S}} \mid l \in [C]\}$, $j \in [d]$.

- $\mathcal{R}$: $\bigcap_{j=1}^{d} W_j$; $\mathcal{S}$: Nothing.

Figure 11: The ideal functionality $\mathcal{F}_{\mathsf{S^3I}}$.

6.3. Instantiating Secret-Shared Set Intersection

We now present our instantiation of $\mathcal{F}_{\mathsf{S^3I}}$ via rOLE. As briefly discussed in Section 2.5, our construction is based on algebraic techniques based on polynomial encodings [23], [24], [28]. While other approaches may be possible thanks to Theorem 3, we stick to such a classical approach for simplicity; it suffices to implement its algebraic representations of set intersection via two-party computation.

Polynomial Encoding of Sets and Their Intersections. For a (multi-)set S of size C, we encode S as a degree-C polynomial $f_S(X) := \prod_{z \in S}(X - z)$. Here, $f_{S_1 \uplus S_2}(X) = f_{S_1}(X) f_{S_2}(X)$ holds for any sets S_1 and S_2. We can decode the set from the encoded polynomial through factorization. With this encoding, we compute the (multiset) intersection of encoded sets via random linear combination as follows.

Theorem 5 ([23, Lemma 2]). *Let $S_1, \ldots, S_d$ be multisets of size C. Then for $S_I = \bigcap_{i=1}^{d} S_i$, the following two distributions are equivalent:*

$$\{\Sigma_{j=1}^{d} \gamma_j(X) f_{S_j}(X) \mid \gamma_j(X) \xleftarrow{\$} \mathbb{F}[X]^{\leq C-1}\},$$
$$\{\gamma(X) \cdot \Pi_{z \in S_I}(X - z) \mid \gamma(X) \xleftarrow{\$} \mathbb{F}[X]^{\leq 2C - |S_I| - 1}\}$$

Proof. See Section B.1 of Kissner and Song [23]. □

Instead of extracting encoded elements from factorizing $\sum_{i=1}^{d} \gamma_j(X) f_{S_j}(X)$, one can utilize the gcd trick proposed by Yeo and Ying [24]. They showed that computing the $\gcd$ of the two independent random linear combination polynomials suffices. Note that Theorem 5 ensures that we can safely replace random polynomial multiplications with two independent random linear combinations.

Theorem 6 (Adapted from [24, Proposition 7]). *Let $\mathbb{F}$ be a finite field and S be a multiset of cardinality $\leq C \in \mathbb{N}$. For S's encoding $f_S(X)$ and independent random polynomials $\gamma_1(X)$, $\gamma_2(X) \in \mathbb{F}[X]^{\leq 2C - |S| - 1}$, we have that*

$$\Pr[\gcd(f_S(X)\gamma_1(X), f_S(X)\gamma_2(X)) \neq f_S(X)] \leq \frac{1}{|\mathbb{F}| - 1}.$$

Here, since Theorem 3 ensures that every element of S_I has multiplicity 1 with high probability, running the zero-finding algorithm suffices instead of the full factorization.

Parameters: Two parties, the receiver $\mathcal{R}$ and the sender $\mathcal{S}$. The key space $\mathcal{K}$ and a field $\mathbb{F}$. The upper bound C of each $\mathcal{R}$'s set. The total cardinality V of $\mathcal{R}$'s sets and the number d of query keys by $\mathcal{S}$.
Inputs:

- $\mathcal{R}$: Q key-set pairs $\{(k_i, S_i)\}_{i=1}^{Q} \subset \mathcal{K} \times \mathcal{P}(\mathbb{F})$ with distinct keys $k_i \in \mathcal{K}$, $|S_i| \leq C$, and $\sum_{i=1}^{Q} |S_i| = V$.
- $\mathcal{S}$: d keys $\hat{k}_1, \ldots, \hat{k}_d$.

Procedure:

(1) $\mathcal{R}$ parses $S_i \to (s_{i,l})_{l=1}^{m_i}$, where $m_i = |S_i|$, and for $V = \sum_{i=1}^{Q} m_i$, set $(k'_q, v'_q) \leftarrow (k_i \| l, s_{i,l})$ s.t. $q = \sum_{i'=1}^{i-1} m_{i'} + l$, $i \in [Q]$, $l \in [m_i]$.
(2) For $T = Cd$, $\mathcal{S}$ sets $\tilde{k}_t \leftarrow \hat{k}_j \| l$, where we denote $t = C \cdot (j-1) + l$ for $j \in [d]$ and $l \in [C]$.
(3) $\mathcal{R}$ and $\mathcal{S}$ invoke $\mathcal{F}_{\mathsf{soOPPRF}}$ on inputs $\{(k'_q, v'_q)\}_{q=1}^{V}$ and $(\tilde{k}_t)_{t=1}^{T}$ by playing the role of $\mathcal{A}$ and $\mathcal{B}$, respectively, obtaining $w_t^{\mathcal{R}}$ and $w_t^{\mathcal{S}}$ for $t \in [T]$.
(4) $\mathcal{R}$ and $\mathcal{S}$ parse $w_t^{\mathcal{R}} \to (w_{j,l}^{\mathcal{R}})_{l=1}^{C}$, $w_t^{\mathcal{S}} \to (w_{j,l}^{\mathcal{S}})_{l=1}^{C}$ for $t = C \cdot (j-1) + l$, $j \in [d]$, $l \in [C]$. Then set $W_j^{\mathcal{R}} \leftarrow (w_{j,l}^{\mathcal{R}})_{l=1}^{C}$ and $W_j^{\mathcal{S}} \leftarrow (w_{j,l}^{\mathcal{S}})_{l=1}^{C}$.
(5) $\mathcal{R}$ and $\mathcal{S}$ invoke $\mathcal{F}_{\mathsf{S^3I}}$ on inputs $(W_j^{\mathcal{R}})_{j=1}^{d}$ and $(W_j^{\mathcal{S}})_{j=1}^{d}$. After, $\mathcal{R}$ obtains a set $W \subset \mathbb{F}$.
(6) $\mathcal{R}$ returns W.

Figure 12: Protocol Π_{OSSI} in $(\mathcal{F}_{\mathsf{soOPPRF}}, \mathcal{F}_{\mathsf{S^3I}})$-hybrid model.

Secret-Shared Encoding. With polynomial encoding, the set intersection can be computed via algebraic operations. We show how to compute $f_S(X)$ via a subset tree with $O(|S| \log |S|)$ multiplications of secret-shared values in $\mathbb{F}$. For simplicity, let $|S| = C = 2^n$ and $S = \{s_1, \ldots, s_{2^n}\}$. We build the following binary tree of depth n:

- Each leaf node holds a singleton $\{s_i\}$, $i \in [2^n]$.
- The parent holds the multiset union of children's sets.

Thus, we can compute $f_S(X)$ by traversing this tree.

Note that each polynomial encoding is monic. For sets S_1, S_2 such that $|S_1| = |S_2| = 2^k$, we can uniquely write $f_{S_1}(X) = X^{2^k} + g_{S_1}(X)$, $f_{S_2}(X) = X^{2^k} + g_{S_2}(X)$ for polynomials $g_{S_1}(X)$ and $g_{S_2}(X)$ of degree $2^k - 1$. Hence, we can rewrite $f_{S_1 \uplus S_2}(X)$ as

$$X^{2^{k+1}} + X^{2^k}(g_{S_1}(X) + g_{S_2}(X)) + g_{S_1}(X) g_{S_2}(X).$$

This gives $g_{S_1 \uplus S_2}(X)$ corresponding to $f_{S_1 \uplus S_2}(X)$:

$$g_{S_1 \uplus S_2}(X) = X^{2^k}(g_{S_1}(X) + g_{S_2}(X)) + g_{S_1}(X) g_{S_2}(X).$$

If $\mathcal{R}$ and $\mathcal{S}$ have a share of $g_{S_1}(X)$ and $g_{S_2}(X)$, they can obtain that of $g_{S_1 \uplus S_2}(X)$ by invoking $\mathcal{F}_{\mathsf{SSrOLE}}^{2^k-1, 2^k-1}$. Since the constant polynomials for leaf nodes correspond to the output of so-OPPRF, i.e., already shared across $\mathcal{R}$ and $\mathcal{S}$, 2^{k-1} invocations of $\mathcal{F}_{\mathsf{SSrOLE}}^{2^{n-k}-1, 2^{n-k}-1}$ suffice for $k \in [n]$. Note that one invocation of $\mathcal{F}_{\mathsf{SSrOLE}}^{2^{n-k}-1, 2^{n-k}-1}$ requires $(2^{n-k+2} - 2)$ OLEs. Thus, computing a share of $f_S(X)$ requires $\sum_{k=1}^{n} 2^{k-1} \cdot (2^{n-k+2} - 2) = n \cdot 2^{n+1} - 2^{n+1} + 2$ OLEs in total, i.e., $2C \log C - 2C + 2$.

Parameters: Two parties, the receiver $\mathcal{R}$ and the sender $\mathcal{S}$. A field $\mathbb{F}$. The number of sets d and cardinality $C \in \mathbb{N}$.
Inputs:

- $\mathcal{R}$: d ordered sets $W_1^{\mathcal{R}}, \ldots, W_d^{\mathcal{R}}$ of cardinality C.
- $\mathcal{S}$: d ordered sets $W_1^{\mathcal{S}}, \ldots, W_d^{\mathcal{S}}$ of cardinality C.

Procedure:

(1) For $\mathcal{P} \in \{\mathcal{R}, \mathcal{S}\}$, parse input sets $W_j^{\mathcal{P}} \to (w_{j,l}^{\mathcal{P}})_{l=1}^{C}$. Then initialize polynomials $g_{j,l}^{\mathcal{P}}(X) \leftarrow -w_{j,l}^{\mathcal{P}}$.

(2) For $l \in [\log C]$ do:

- For $\hat{l} \in [\frac{C}{2^l}]$, $\mathcal{R}$ and $\mathcal{S}$ invoke $\mathcal{F}_{\mathsf{SSrOLE}}^{2^{l-1}-1,2^{l-1}-1}$ on input polynomials $(g_{j,2\hat{l}-1}^{\mathcal{R}}(X), g_{j,2\hat{l}}^{\mathcal{R}}(X))$ and $(g_{j,2\hat{l}-1}^{\mathcal{S}}(X), g_{j,2\hat{l}}^{\mathcal{S}}(X))$, obtaining $\hat{g}_{j,\hat{l}}^{\mathcal{R}}(X)$, $\hat{g}_{j,\hat{l}}^{\mathcal{S}}(X)$.
- For $\hat{l} \in [\frac{C}{2^l}]$ and $\mathcal{P} \in \{\mathcal{R}, \mathcal{S}\}$, update:

$$g_{j,\hat{l}}^{\mathcal{P}}(X) \leftarrow X^{2^{l-1}}(g_{j,2\hat{l}-1}^{\mathcal{P}}(X) + g_{j,2\hat{l}}^{\mathcal{P}}(X)) + \hat{g}_{j,\hat{l}}^{\mathcal{P}}(X)$$

(3) For $j \in [d]$ and $\mathcal{P} \in \{\mathcal{R}, \mathcal{S}\}$, set $g_j^{\mathcal{P}}(X) \leftarrow g_{j,1}^{\mathcal{P}}(X)$.

(4) For $j \in [d]$, $\mathcal{S}$ sets $\gamma_{j,1}^{\mathcal{S}}(X), \gamma_{j,2}^{\mathcal{S}}(X) \xleftarrow{\$} \mathbb{F}[X]^{\leq C-1}$, and $\mathcal{R}$ and $\mathcal{S}$ invoke $\mathcal{F}_{\mathsf{rOLE}}^{C-1,C-1}$ twice on $g_j^{\mathcal{R}}(X)$ and $\gamma_{j,i}^{\mathcal{S}}(X)$, $i \in \{1,2\}$, obtaining $h_{j,i}^{\mathcal{R}}(X)$ and $h_{j,i}^{\mathcal{S}}(X)$.

(5) For $i \in \{1,2\}$, $\mathcal{S}$ computes

$$h_i^{\mathcal{S}}(X) \leftarrow \sum_{j=1}^{d} (\gamma_{j,i}^{\mathcal{S}}(X)X^C + h_{j,i}^{\mathcal{S}}(X) + \gamma_{j,i}^{\mathcal{S}}(X) g_j^{\mathcal{S}}(X))$$

and send $(h_1^{\mathcal{S}}(X), h_2^{\mathcal{S}}(X))$ to $\mathcal{R}$.

(6) $\mathcal{R}$ sets $h_i^{\mathcal{R}}(X) \leftarrow \sum_{j=1}^{d} h_{j,i}^{\mathcal{R}}(X)$, $i \in \{1,2\}$, and

$$A(X) \leftarrow \gcd(h_1^{\mathcal{R}}(X) + h_1^{\mathcal{S}}(X), h_2^{\mathcal{R}}(X) + h_2^{\mathcal{S}}(X))$$

(7) $\mathcal{R}$ returns $\{z \in \mathbb{F} \mid A(z) = 0\}$.

Figure 13: Protocol $\Pi_{\mathsf{S^3I}}$ in $\mathcal{F}_{\mathsf{SSrOLE}}$-hybrid model.

Instantiating $\mathcal{F}_{\mathsf{S^3I}}$. With the above idea, we can compute the share of the encoding polynomial. The random linear combination can be done via $\mathcal{F}_{\mathrm{rOLE}}^{C-1,C-1}$ since $\gamma(X)f_S(X) = \gamma(X)X^C + \gamma(X)g_S(X)$, i.e., a rOLE for $\gamma(X)g_S(X)$ suffices. The gcd trick can be used after revealing $\mathcal{S}$'s share of final polynomials to $\mathcal{R}$. Thus, we can instantiate $\mathcal{F}_{\mathsf{S^3I}}$ in $\mathcal{F}_{\mathsf{OLE}}$-hybrid model, as $\Pi_{\mathsf{S^3I}}$ described in Fig. 13.

Theorem 7. *$\Pi_{\mathsf{S^3I}}$ in Fig. 13 securely implements $\mathcal{F}_{\mathsf{S^3I}}$ against semi-honest adversaries in $\mathcal{F}_{\mathsf{SSrOLE}}$-hybrid model.*

Remark. Since our construction uses $\mathcal{F}_{\mathsf{SSrOLE}}$ as a black box, it can be instantiated with any protocol realizing this functionality; for instance, the ring OLE of Chongchitmate et al. [28] can be used for this purpose. To demonstrate the applicability of our approach in the simplest form, we describe our construction using plain OLE (cf. Appendix A).

6.4. Putting Them All Together

By replacing $\mathcal{F}_{\mathsf{S^3I}}$ in Π_{OSSI} with $\Pi_{\mathsf{S^3I}}$, we can securely implement $\mathcal{F}_{\mathsf{OSSI}}$ in $(\mathcal{F}_{\mathsf{soOPPRF}}, \mathcal{F}_{\mathsf{OLE}})$-hybrid model. Hence, from these building blocks, we obtain the following result.

Theorem 8. *Π_{FPSI} (Fig. 8) instantiated with Π_{OSSI} (Fig. 12) and $\Pi_{\mathsf{S^3I}}$ (Fig. 13) securely implements $\mathcal{F}_{\mathsf{FPSI}}$ against semi-honest adversaries in $(\mathcal{F}_{\mathsf{soOPPRF}}, \mathcal{F}_{\mathsf{OLE}}, \mathcal{F}_{\mathsf{OT}})$-hybrid model.*

Proof. By Theorem 4 and Theorem 7, we know that Π_{OSSI} instantiated with $\Pi_{\mathsf{S^3I}}$ securely implements $\mathcal{F}_{\mathsf{OSSI}}$ against semi-honest adversaries in $(\mathcal{F}_{\mathsf{soOPPRF}}, \mathcal{F}_{\mathsf{OLE}})$-hybrid model. Hence, by Theorem 2, we obtain the desired result. □

Optimization with Skipping the Factorization. We can optimize Π_{FPSI} by replacing $\Pi_{\mathsf{S^3I}}$'s final step in Π_{OSSI} by letting $\mathcal{R}$ check whether the degree of the gcd polynomial is nonzero or not. This change does not affect $\mathcal{R}$'s view in the FPSI protocol, as it retrieves $\mathcal{S}$'s matched items, while avoiding the need for factorization of the gcd polynomial.

A Remark on the Field Size $|\mathbb{F}|$. As we can observe from Theorem 3 and 6, one invocation of $\mathcal{F}_{\mathsf{OSSI}}$ involves an error probability of $\frac{C^2}{|\mathbb{F}|} + \frac{C^{d+1}}{|\mathbb{F}|^d} + \frac{1}{|\mathbb{F}|-1}$. Since Π_{FPSI} invokes $\mathcal{F}_{\mathsf{OSSI}}$ N times, by the union bound, we obtain the total error probability bound $O(\frac{NC^2}{|\mathbb{F}|})$. This implies that, to ensure κ-bit statistical security, we must select $|\mathbb{F}| \geq NC^2 \cdot 2^\kappa$.

6.5. Efficiency Analysis

Throughout the protocol Π_{FPSI}, we summarize the invocations of each subroutine $\mathcal{F}_{\mathsf{soOPPRF}}, \mathcal{F}_{\mathsf{OLE}}, \mathcal{F}_{\mathsf{OT}}$ as follows.

- $\mathcal{F}_{\mathsf{soOPPRF}}$: $\mathcal{R}$ programs $(2\delta+1)dM$ input-output pairs as each item of $\mathcal{R}$ contributes $(2\delta+1)d$ pairs. $\mathcal{S}$ queries dCN inputs. It requires communicating $O(\delta dM + dCN)$ field elements across 2 rounds, and $\mathcal{R}$ and $\mathcal{S}$ require $O(\delta dM + dCN)$ and $O(dCN)$ computations, respectively.
- $\mathcal{F}_{\mathsf{OLE}}$: $\mathcal{R}$ and $\mathcal{S}$ invoke $O(dCN \log C)$ OLEs, incurring $O(dCN \log C)$ computations on both parties, and $O(dCN \log C)$ communication in field elements across $\log C$ rounds.
- $\mathcal{F}_{\mathsf{OT}}$: $\mathcal{R}$ and $\mathcal{S}$ incur N OTs to retrieve elements in $\mathcal{U}^d$ across 1 round with $O(udN)$ bits of communication.

In addition to these computation costs, we also consider additional computations during $\Pi_{\mathsf{S^3I}}$, including (i) sending $(h_1^{\mathcal{S}}, h_2^{\mathcal{S}})$, and (ii) computing gcd. The former takes $O(CN)$ of communication, and the latter takes $O(C^2N)$ of computation. To sum up, each party's computation/communication costs and round complexity are as follows:

- $\mathcal{R}$'s Computation: $O(\delta dM + dCN \log C + C^2N)$
- $\mathcal{S}$'s Computation: $O(dCN \log C)$
- Communication & Rounds: $O(\delta dM + dCN \log C)$ of communication across $(4 + \log C)$ rounds.

For example, our protocol for uniformly distributed data where $C = O(\frac{\log d}{\log \log d})$, exhibits $O(\delta dM + dN \log d)$ computation and communication complexities, i.e., quasi-linear in d and linear in M and N.

7. Experimental Analysis

All experiments were conducted on a machine with an Intel Xeon Gold 6448H CPU (core clock of 2.40 GHz) and 512 GB of RAM, in a single-threaded setting.

7.1. Concrete Analysis on C-Max Collision

We first analyze the plausibility of the C-max collision property under the uniform distribution with various data parameter settings (d, M, δ) and real-world datasets.

Empirical C Values Uniform Distribution. We first i.i.d. sample data from the uniform distribution over $\mathcal{U}$, and measure the exact C value according to Lemma 2. We sweep $\delta \in \{10, 30, 60\}$, $M \in \{2^8, 2^{12}, 2^{16}, 2^{20}\}$, and $d \in \{2, 6, 10, 16\}$. We fix $u = 32$ for the universe $\mathcal{U}$. We report the average and maximum C values across 1,000 trials, and compare the measured C values with the conservative ones estimated from Corollary 1. Tab. 1 reports the result. Empirically measured C values are less than half of the corresponding conservative bounds on average, and they gracefully increase as M, d, and δ grow. The observed C is at most 6 in all settings. We guess that such a gap between empirical values and the theoretical bounds comes from the looseness of the union bound over $\mathcal{U}$ in the proof of Theorem 1.

Evaluations on Real-World Datasets. In addition to C values for uniformly distributed data, we also evaluate the concrete values on public real-world datasets: COVID locations of interest data in New Zealand[3] (NZ-COVID) and enterprise authentication log data from Los Alamos National Laboratory[4] (LANL). NZ-COVID contains coordinates of locations where COVID patients have visited in New Zealand, and LANL records the timestamp when an authentication happens. The former corresponds to location-based contact tracing, while the latter corresponds to bots/duplicated users detection via their authentication patterns in a fixed time window. After preprocessing, we obtain $(M, d) = (2167, 2)$ for NZ-COVID and $(1026, 24)$ for LANL, representing settings for low- and high-dimensional data, respectively. Details of these datasets are provided in Appendix F.

For NZ-COVID, we obtain $C = 14$ and 30 for $\delta = 20$m and $\delta = 50$m, respectively. For LANL, we obtain $C = 11$, 13, and 21 for $\delta = 10$s, 30s, and 60s, respectively. Because these datasets are non-uniformly distributed, the obtained C values are consistently larger than those from uniform data. Nevertheless, these values are substantially smaller than the total number of items, 2,167 and 1,026 for NZ-COVID and LANL, respectively, suggesting that the proposed C-max collision assumption is non-vacuous and can hold with C far below M on real data.

7.2. Benchmarking the Proposed FPSI Protocol

We benchmark our protocol on various parameters. We consider the LAN network—10 Gbps bandwidth and 0.2 ms RTT—and estimate the communication latency. Following prior FPSI works [3], [4], [9], [11], [16], [19], [29], we primarily focus on online costs. Offline costs, e.g., local computation or preprocessing, are reported in Appendix E.2.

3. https://github.com/minhealthnz/nz-covid-data
4. https://csr.lanl.gov/data/cyber1/

TABLE 1: The measured average/maximum C values, and the estimated ones (in parentheses) from Corollary 1.

M	δ	$d=2$	$d=6$	$d=10$	$d=16$
2^8	10	1.0/2 (4)	1.0/2 (4)	1.0/2 (4)	1.0/2 (4)
	30	1.0/1 (4)	1.0/2 (5)	1.0/2 (5)	1.0/2 (5)
	60	1.0/2 (5)	1.0/2 (5)	1.0/2 (5)	1.0/2 (5)
2^{12}	10	1.2/2 (5)	1.3/2 (5)	1.5/2 (5)	1.7/2 (5)
	30	1.4/2 (5)	1.7/2 (5)	1.9/2 (6)	2.0/2 (6)
	60	1.6/2 (6)	1.9/2 (6)	2.0/2 (6)	2.0/2 (6)
2^{16}	10	2.0/3 (6)	2.0/3 (6)	2.0/3 (6)	2.0/3 (6)
	30	2.1/3 (7)	2.1/3 (7)	2.2/3 (7)	2.4/3 (7)
	60	2.2/3 (7)	2.5/4 (8)	2.7/4 (8)	2.8/4 (8)
2^{20}	10	3.0/4 (8)	3.1/4 (9)	3.2/4 (9)	3.3/4 (9)
	30	3.7/5 (10)	4.0/5 (10)	4.0/5 (10)	4.0/6 (10)
	60	4.1/5 (11)	4.2/5 (11)	4.3/6 (11)	4.4/6 (11)

Implementation Details. We implement our protocol in C/C++, instantiating core building blocks as follows:

- We implement $\mathcal{F}_{\mathsf{soOPPRF}}$ via Alamati et al.'s $(\mathbb{F}_3, \mathbb{F}_2)$ alternating moduli PRF [22][5] and Bienstock et al.'s RB-OKVS [30]. We regard the PRF output space as $\mathbb{F}_{2^{64}}$ by truncating the original 128-bit output.
- We utilize PCG for $\mathcal{F}_{\mathsf{OLE}}$ based on Ring-LPN with a distributed multipoint function [31][6].
- We utilize random OT for $\mathcal{F}_{\mathsf{OT}}$ based on subfield VOLE implemented in libOTe library [7].

We utilize AVX-512 SIMD instructions and `PCLMULQDQ` instructions to optimize field operations over $\mathbb{F}_{2^{64}}$, ensuring 128-bit computational security and 40-bit statistical security in all experiments. We provide detailed micro-optimizations in Appendix E.1. Remark that our instantiation allows offline/local computations, e.g., local OKVS encoding by $\mathcal{R}$ or OLE/OT correlation generation.

Evaluation Result. Based on our analyses in Section 7.1, we select three C values, $4, 8, 16$, corresponding to empirically observed, theoretically guaranteed, and conservative ones in the uniform distribution. We evaluate the online latency and the communication cost on various settings, including set sizes $M = N \in \{2^8, 2^{12}, 2^{16}\}$, the threshold $\delta \in \{10, 30\}$, and the data dimensionality $d \in \{2, 6, 10, 16\}$. This reflects parameter settings of separation-based FPSI protocols for low-dimensional [6], [7], [8], [14] ($d \leq 5$) and high-dimensional regimes [9], [11], [16], [18] ($d \geq 6$).

The results are provided in Tab. 2. We observe that the threshold δ does not significantly affect the online latency, as it primarily affects the OKVS encoding, i.e., the local computation. Rather, it affects the communication cost for sending an OKVS when invoking so-OPPRF. Meanwhile, for the set size and the dimension, both computation and communication costs linearly increase for each fixed C. For C, the cost increases slightly faster than linear due to the $O(C \log C)$ factor on the number of OLE correlations. These results provide insights into how our protocol's cost scales with real-world datasets according to the parameters we obtained in Section 7.1. We provide a detailed cost breakdown in Appendix E.3.

5. https://github.com/ladnir/secure-join
6. https://github.com/MatanHamilis/dmpf
7. https://github.com/osu-crypto/libOTe

TABLE 2: Benchmark results of the proposed protocol in communication cost (C.; MB) and online latency (T.; s).

$M{=}N$	C	(2,10)		(2,30)		(6,10)		(6,30)		(10,10)		(10,30)		(16,10)		(16,30)	
		C.	T.	C.	T.	C.	T.	C.	T.	C.	T.	C.	T.	C.	T.	C.	T.
2^8	4	0.607	0.001	0.784	0.002	1.75	0.004	2.28	0.004	2.90	0.006	3.78	0.007	4.61	0.010	6.03	0.012
	8	1.23	0.003	1.41	0.003	3.55	0.009	4.08	0.009	5.88	0.014	6.76	0.014	9.36	0.024	10.78	0.023
	16	2.61	0.007	2.78	0.007	7.55	0.021	8.08	0.019	12.50	0.035	13.38	0.032	19.91	0.055	21.33	0.052
2^{12}	4	9.68	0.022	12.51	0.025	27.99	0.068	36.48	0.086	46.30	0.123	60.45	0.142	73.76	0.208	96.41	0.231
	8	19.64	0.050	22.47	0.052	56.82	0.153	65.32	0.171	94.01	0.263	108.2	0.283	149.8	0.440	172.4	0.473
	16	41.66	0.119	44.49	0.117	120.8	0.360	129.3	0.365	199.9	0.604	214.1	0.611	318.6	1.01	341.3	0.998
2^{16}	4	154.9	0.445	200.2	0.505	447.8	1.35	583.7	1.56	740.7	2.34	967.2	2.63	1180	3.88	1542	4.25
	8	314.2	0.925	359.5	1.02	909.2	2.85	1045	3.05	1504	4.92	1731	5.18	2396	8.06	2759	8.42
	16	666.6	2.16	711.9	2.20	1933	6.45	2068	6.73	3199	11.26	3425	12.22	5098	18.08	5460	18.54

TABLE 3: Communication (C.; MB) and online latency (T.; s) of ours relative to state-of-the-art separation-based FPSI protocols for $M = N = 2^{16}$. †: Their assumption is not well-justified. ‡: extrapolated according to their complexity 2^d.

Protocols (d, δ)	Assump.	(2,10)		(2,30)		(6,10)		(6,30)		(10,10)		(10,30)		(16,10)		(16,30)	
		C.	T.	C.	T.	C.	T.	C.	T.	C.	T.	C.	T.	C.	T.	C.	T.
Piske et al. [14]	$\mathcal{S}$ disj. hash.	34.24	38.29	34.24	36.97	612.9	215.0	612.9	218.5	13906	3000	13906	3029	$1.4{\times}10^{6‡}$	$3.1{\times}10^{5‡}$	$1.4{\times}10^{6‡}$	$3.1{\times}10^{5‡}$
Yang et al. [16]	$\mathcal{R} \wedge \mathcal{S}$ disj. proj.	363.1†	2.13†	705.3†	5.02†	927.7	5.64	1960	14.30	1495	9.21	3213	23.67	2345	14.71	5087	38.81
Ours ($C = 4$)	$\mathcal{R}$ C-max coll.	154.9	0.45	200.2	0.50	447.8	1.35	583.7	1.56	740.7	2.34	967.2	2.63	1180	3.88	1542	4.25
Ours ($C = 8$)		314.2	0.93	359.5	1.02	909.2	2.85	1045	3.05	1504	4.92	1731	5.18	2396	8.06	2759	8.42
Ours ($C = 16$)		666.6	2.16	711.9	2.20	1933	6.45	2068	6.73	3199	11.26	3425	12.22	5098	18.08	5460	18.54

7.3. Reference Separation-based FPSI Protocols

We evaluate two state-of-the-art FPSI protocols: Piske et al. [14] based on spatial hashing, and Yang et al. [16] based on disjoint projection. We evaluate them for the ℓ_∞ metric. While a direct apples-to-apples comparison is not possible, they serve as a reference point to show the concrete efficiency of our protocol. With $M{=}N{=}2^{16}$, we evaluate the online latency and communication cost under the parameters in Section 7.2; comparison on the end-to-end cost including offline costs is provided in Appendix E.2. All protocols ensure 128-bit computational and 40-bit statistical security.

Tab. 3 presents the results. Across all settings, our protocol exhibits online latency comparable to or lower than that of the other protocols. Ours offloads all heavy procedures—$\mathcal{R}$'s OKVS encoding in so-OPPRF and OLE/OT correlation generation—to offline/local computations, and Yang et al. [16] likewise count some of their OKVS encodings as offline. While these protocols can also utilize silent OLE/OT preprocessing[8], unlike ours, they still require some OKVS encodings online. Compared to Piske et al. [14], which favors the low-dimensional case $d = 2$, ours incurs a larger communication cost. However, their overheads drastically increase as d grows, and the crossover occurs when $d = 10$. In contrast, Yang et al. [16] share the same complexity regime as ours with a fixed C. When $C = 16$, their protocol incurs lower communication cost than ours; for $C \leq 8$, however, ours achieves comparable or lower communication cost. We note that the C values measured on the real-world datasets (Section 7.1) range from 11 to 30, and thus fall outside the $C \leq 8$ regime. However, the disjoint-projection assumption required by Yang et al. fails on both datasets, so theirs does not apply to either dataset.

8. Discussion and Limitation

Extended Functionality, Metrics, and Security. Though we focused on the functionality to reveal $\mathcal{S}$'s item, other variants have also considered in the literature, such as labeled FPSI [4], FPSI with cardinality [7], or FPSI with sender privacy [8], i.e., letting $\mathcal{R}$ learn its own matched item without revealing $\mathcal{S}$'s item. To achieve this, we may consider a variant of OSSI that outputs a secret-shared bit indicating whether the intersection is nonempty; given such a bit, these functionalities follow in the circuit-based PSI paradigm [32]. However, an efficient instantiation of such a variant remains open, though a straightforward approach is to compute $\gcd$ in $\Pi_{\mathsf{S^3I}}$ in secret-shared form and test whether any of its non-constant coefficients is nonzero.

Extending our protocol to general ℓ_p distances is also possible by representing p-powered ℓ_p norm as an axis-wise summation [8], [16]. This requires a secret-shared variant of the PSI-SUM functionality [33] in place of $\mathcal{F}_{\mathsf{S^3I}}$, followed by a threshold comparison on secret-shared p-powered ℓ_p norm. One caveat is that handling the p-powered ℓ_p norm requires a field of sufficiently large characteristic, so share conversion is necessary before $\Pi_{\mathsf{S^3I}}$. Designing a density-bounded assumption for Hamming distance is non-trivial because each component takes only a bit; seeking an assumption analogous to the UniqC assumption by Gao et al. [9] would also be an interesting future direction.

Like almost all prior FPSI studies [3], [13], [14], [15], [16], [19], [34], we focus on the semi-honest model, and extending it to achieve malicious security is also an important future direction. One may employ a recent maliciously secure OPRF [35] based on alternating moduli, with their maliciously secure VOLE construction. However, as pointed out by Yang et al. [16], verifying the well-formedness of OKVS remains the key technical challenge.

Limitations. While our protocol achieves low online latency, it incurs heavy local/offline computations, e.g., OT/OLE correlation generation and local OKVS encoding by $\mathcal{R}$. Nevertheless, the OKVS encoding can be reused when $\mathcal{R}$'s set remains fixed across multiple executions; $\mathcal{S}$ never sees the actual PRF evaluation result during the protocol and thus learns nothing from a reused encoding. Moreover, since PCG generation is input-independent, both parties

8. Their implementation generates correlations on the fly; we therefore modified the code to exclude correlation generation from the online latency.

can preprocess their OT/OLE correlation in advance when they are not engaged in the FPSI protocol. We also note that, as with the disjoint projection assumption, the C-max collision property depends heavily on the data distribution; in the worst case when $C = M$, the protocol's computation and communication complexities become $O(M^2N)$ and $O(MN \log M)$, respectively. However, the key difference is that ours degrades gracefully: a denser distribution can be handled simply by raising the C budget with a corresponding increase in cost, rather than by a structural failure.

9. Related Work

Spatial Hashing. Spatial hashing has been a central building block of FPSI protocols over the Euclidean space [6], [7], [8], [10], [11], [14], [15], [17], [19], [34]. This technique allows for filtering out non-matching items via hashed values, effectively reducing the number of candidates for closeness matching. While the protocol by Garimella et al. [6]—the first FPSI construction based on spatial hashing—was for ℓ_∞ metric, subsequent studies, including van Baarsen and Pu [8], [15], Richardson et al. [10], and Piske et al. [14], extend it for general ℓ_p distances. However, spatial hashing-based constructions exhibit $\Omega(2^d)$ communication and/or computation complexity, which is inherited from the matching ambiguity of 2^d candidates. In addition, these protocols typically require that all set items are far apart by at least 2δ or more for the decision threshold δ—so-called "apart by 2δ" assumption. Structural assumptions regarding spatial hashing have also been proposed, such as the disjoint hash [10], [14] or disjoint mini-universe [7], [34] assumptions, where spatial hashing values or grid cell coordinates do not overlap across set items.

Disjoint Projection Assumption. To mitigate $\Omega(2^d)$ overhead, several works study additional structural assumptions stronger than "apart by $> 2\delta$". Among them, the disjoint projection assumption showed the feasibility of designing an efficient FPSI with $O(\text{poly}(d))$ overhead [6], [8], [9], [15]. The disjoint projection assumption requires that, for every item x, there exists an axis j such that no other item lies within distance 2δ of x along the jth coordinate.

This assumption was first studied by Garimella et al. [6], who propose the first FPSI protocol for ℓ_∞ metric with $O(d^2)$ overhead. The construction of Garimella et al. relies on a stronger form: for every item and axis, no other item lies within distance 2δ along the axis—called the global disjoint projection. Later, van Baarsen and Pu [8] relaxed their assumption to the disjoint projection assumption as stated above, maintaining the same $O(d^2)$ overhead. Gao et al. [9] further improved this complexity to $O(d)$, proposing other constructions for ℓ_1, ℓ_2, and Hamming metrics, and Yang et al. [16] later improved its concrete efficiency via symmetric key primitives. Disjoint projection is often justified for high-dimensional data. van Baarsen and Pu proved that, for the uniformly sampled data points, the disjoint projection assumption holds with a probability of $1-\mathsf{negl}(d)$ for some negligible function negl [8, Lemma 7].

Constructions from Other Approaches. There exist FPSI proposals that do not rely on separation-based assumptions and spatial hashing [3], [4], [5], [13], [29], [36], [37]. They often directly model similarity matching (or its efficient approximation) as a circuit, and employ secure computation techniques for evaluating such a circuit, e.g., (V)OLE-based computation [3], [29], [37] or (fully) homomorphic encryption [4], [5]. While these constructions do not require structural assumptions on the data in general, they require $O(MN)$ computation and/or communication complexity overhead derived from pairwise similarity matching [3], [4], [5], [13], [29], [37]. Although some constructions showed that a subquadratic communication/computation cost is achievable [36], they require another variant of separation-based assumption to avoid the quadratic blow-up. Recently, van Baarsen and Pu [15] proposed an FPSI protocol without relying on any assumptions. While their protocol achieves linear complexity in both M and N, it exhibits $O((\log \delta)^d)$ overhead. A concurrent work by Singh et al. [12] also considered the density-bounded assumptions, but they inherit the same limitation of separation-based FPSI protocols as they use those protocols as a subroutine.

10. Conclusion

In this study, we present an efficient FPSI protocol based on the new density-bounded assumption, C-max collision. With our membership tag-based abstraction and its instantiation based on the novel OSSI and $\mathsf{S^3I}$ functionality, we design the first efficient FPSI protocol for the ℓ_∞ metric that effectively supports both low- and high-dimensional data simultaneously. Experimentally, our protocol shows concrete online efficiency comparable to state-of-the-art separation-based protocols. We believe newly proposed functionalities are of independent interest and may be useful for privacy-preserving computations beyond FPSI. Our result demonstrates that prevailing separation-based paradigms may not be fundamental for achieving efficient FPSI, and we hope that our approach opens new directions for FPSI protocol design. Extending our approach to general ℓ_p metrics and richer functionalities remains as interesting future work.

Acknowledgments

The authors thank the anonymous reviewers and shepherd for their constructive comments and feedback. This work was supported in part by the Institute for Information and Communications Technology Planning and Evaluation (IITP), grant funded by the Ministry of Science and ICT (MSIT), Republic of Korea (RS-2024-00399491), in part by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00332210).

References

[1] T. Duong, D. H. Phan, and N. Trieu, "Catalic: Delegated psi cardinality with applications to contact tracing," in *IACR ASIACRYPT*. Springer, 2020, pp. 870–899.

[2] P. Hallgren, C. Orlandi, and A. Sabelfeld, "Privatepool: Privacy-preserving ridesharing," in *IEEE CSF*. IEEE, 2017, pp. 276–291.

[3] T. Rausch, S. Chatel, and W. Lueks, "xdup: Privacy-preserving deduplication for humanitarian organizations using fuzzy psi," *IEEE S&P*, 2026.

[4] E. Uzun, S. P. Chung, V. Kolesnikov, A. Boldyreva, and W. Lee, "Fuzzy labeled private set intersection with applications to private real-time biometric search," in *USENIX Security*, 2021, pp. 911–928.

[5] H. Son, S. Paik, Y. Kim, S. Kim, H. Chung, and J. H. Seo, "Doubly efficient fuzzy private set intersection for high-dimensional data with cosine similarity," *IEEE Access*, vol. 13, pp. 217 108–217 125, 2025.

[6] G. Garimella, M. Rosulek, and J. Singh, "Structure-aware private set intersection, with applications to fuzzy matching," in *IACR CRYPTO*. Springer, 2022, pp. 323–352.

[7] G. Garimella, B. Goff, and P. Miao, "Computation efficient structure-aware psi from incremental function secret sharing," in *IACR CRYPTO*. Springer, 2024, pp. 309–345.

[8] A. van Baarsen and S. Pu, "Fuzzy private set intersection with large hyperballs," in *IACR EUROCRYPT*. Springer, 2024, pp. 340–369.

[9] Y. Gao, L. Qi, X. Liu, Y. Luo, and L. Wang, "Efficient fuzzy private set intersection from fuzzy mapping," in *IACR ASIACRYPT*. Springer, 2024, pp. 36–68.

[10] D. Richardson, M. Rosulek, and J. Xu, "Fuzzy psi via oblivious protocol routing," *Cryptology ePrint Archive*, 2024.

[11] C. Dang, X. Zhou, and B. Liang, "Efficient fuzzy psi based on prefix representation," in *ACM CCS*, 2025, pp. 2204–2218.

[12] S. Singh, Y. Jia, and A. Kate, "Fuzzy private set intersection for real-world datasets," *Cryptology ePrint Archive*, 2026.

[13] E.-O. Blass and G. Noubir, "Assumption-free fuzzy psi via predicate encryption," in *USENIX Security*, 2026, pp. 3003–3022.

[14] L. Piske, J. Singh, N. Trieu, V. Kolesnikov, and V. Zikas, "Distance-aware ot with application to fuzzy psi," in *ACM CCS*, 2025, pp. 4679–4691.

[15] A. van Baarsen and S. Pu, "Fuzzy private set intersection from vole," in *IACR ASIACRYPT*. Springer, 2025, pp. 327–360.

[16] X. Yang, M. Hao, C. Weng, R. H. Deng, Y. Wen, and T. Zhang, "Efficient fuzzy private set intersection from secret-shared oprf," *IEEE S&P*, 2026.

[17] G. Garimella, M. Rosulek, and J. Singh, "Malicious secure, structure-aware private set intersection," in *IACR CRYPTO*. Springer, 2023, pp. 577–610.

[18] C. Zhang, Y. Chen, Y. Cao, Y. Bai, S. Li, J. Lin, A. Wang, and X. Wang, "Fast fuzzy psi from symmetric-key techniques," *Cryptology ePrint Archive*, 2025.

[19] S. Meng, X. Wang, X. Zhou, and B. Liang, "Unbalanced fuzzy private set intersection for l_∞ distance: Achieving sublinear communication with large set size," in *USENIX Security*, 2026, pp. 3299–3318.

[20] V. Kolesnikov, N. Matania, B. Pinkas, M. Rosulek, and N. Trieu, "Practical multi-party private set intersection from symmetric-key techniques," in *ACM CCS*, 2017, pp. 1257–1272.

[21] G. Garimella, B. Pinkas, M. Rosulek, N. Trieu, and A. Yanai, "Oblivious key-value stores and amplification for private set intersection," in *IACR CRYPTO*. Springer, 2021, pp. 395–425.

[22] N. Alamati, G.-V. Policharla, S. Raghuraman, and P. Rindal, "Improved alternating-moduli prfs and post-quantum signatures," in *IACR CRYPTO*. Springer, 2024, pp. 274–308.

[23] L. Kissner and D. Song, "Privacy-preserving set operations," in *IACR CRYPTO*. Springer, 2005, pp. 241–257.

[24] F. Y. Yeo and J. H. Ying, "Private set intersection and other set operations in the third party setting," in *USENIX Security*, 2025, pp. 7723–7742.

[25] Y. Lindell, "How to simulate it–a tutorial on the simulation proof technique," *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*, pp. 277–346, 2017.

[26] R. Motwani and P. Raghavan, "Randomized algorithms," *ACM Computing Surveys (CSUR)*, vol. 28, no. 1, pp. 33–37, 1996.

[27] M. Naor and B. Pinkas, "Oblivious transfer and polynomial evaluation," in *ACM STOC*, 1999, pp. 245–254.

[28] W. Chongchitmate, Y. Ishai, S. Lu, and R. Ostrovsky, "Psi from ring-ole," in *ACM CCS*, 2022, pp. 531–545.

[29] A. Chakraborti, G. Fanti, and M. K. Reiter, "Distance-aware private set intersection," in *USENIX Security*, 2023, pp. 319–336.

[30] A. Bienstock, S. Patel, J. Y. Seo, and K. Yeo, "Near-optimal oblivious key-value stores for efficient psi, psu and volume-hiding multi-maps," in *USENIX Security*, 2023, pp. 301–318.

[31] E. Boyle, N. Gilboa, M. Hamilis, Y. Ishai, and Y. Tu, "Improved constructions for distributed multi-point functions," in *IEEE S&P*. IEEE, 2025, pp. 2414–2432.

[32] B. Pinkas, T. Schneider, O. Tkachenko, and A. Yanai, "Efficient circuit-based psi with linear communication," in *IACR EUROCRYPT*. Springer, 2019, pp. 122–153.

[33] M. Ion, B. Kreuter, A. E. Nergiz, S. Patel, S. Saxena, K. Seth, M. Raykova, D. Shanahan, and M. Yung, "On deploying secure computing: Private intersection-sum-with-cardinality," in *IEEE EuroS&P*. IEEE, 2020, pp. 370–389.

[34] D. Bui, G. Garimella, P. Miao, and P. V. L. Pham, "New framework for structure-aware psi from distributed function secret sharing," in *IACR ASIACRYPT*. Springer, 2025, pp. 294–326.

[35] D. F. Aranha, A. van Baarsen, A. B. Hansen, K. Nielsen, and P. Scholl, "A maliciously-secure post-quantum oprf from crypto dark matter," *IEEE S&P*, 2026.

[36] W. Chongchitmate, S. Lu, and R. Ostrovsky, "Approximate psi with near-linear communication," *Cryptology ePrint Archive*, 2024.

[37] D. Bui and K. Cong, "Efficient fuzzy labeled psi from vector ring-ole," in *CANS*. Springer, 2025, pp. 173–198.

[38] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth, "On the lambert w function," *Advances in Computational Mathematics*, vol. 5, no. 1, pp. 329–359, 1996.

[39] D. J. Bernstein and B.-Y. Yang, "Fast constant-time gcd computation and modular inversion," *IACR TCHES*, pp. 340–398, 2019.

[40] J. P. Snyder, *Map projections: A working manual*. CUP Archive, 1987, vol. 1395.

Appendix A.
Building (SS)rOLE from OLE

For a set of points $\{u_i\}_{i=1}^m, \{z_i\}_{i=1}^m \subset \mathbb{F}$ and a polynomial $f(X) \in \mathbb{F}[X]$, we denote the following two algorithms for multipoint evaluation and Lagrange interpolation.

- $\mathsf{MEval}(f, \{u_i\}_{i=1}^m) \to (f(u_i))_{i=1}^m$.
- $\mathsf{Interp}(\{u_i\}_{i=1}^m, \{z_i\}_{i=1}^m) \to g(X) \in \mathbb{F}[X]^{\leq m-1}$ such that $g(u_i) = z_i$, $\forall i \in [m]$.

With these notations, we show the instantiation $\Pi_{\mathsf{SSrOLE}}^{m,n}$ of $\mathcal{F}_{\mathsf{SSrOLE}}^{m,n}$ under $\mathcal{F}_{\mathsf{OLE}}$-hybrid model in Fig. 14.

We can show that $\Pi_{\mathsf{SSrOLE}}^{m,n}$ securely implements $\mathcal{F}_{\mathsf{SSrOLE}}^{m,n}$ as MEval and Interp constitute an isomorphism between $\mathbb{F}^m$ and $\mathbb{F}[X]^{\leq m-1}$. In Step 4 of $\Pi_{\mathsf{SSrOLE}}^{m,n}$, $\mathcal{F}_{\mathsf{OLE}}$ ensure that

$$\begin{aligned} w_l^{\mathcal{R}} + w_l^{\mathcal{S}} &= z_{1,l}^{\mathcal{R}} z_{2,l}^{\mathcal{R}} + z_{1,l}^{\mathcal{R}} z_{2,l}^{\mathcal{S}} + z_{2,l}^{\mathcal{R}} z_{1,l}^{\mathcal{S}} + z_{1,l}^{\mathcal{S}} z_{2,l}^{\mathcal{S}} \\ &= (f_1^{\mathcal{R}}(u_l) + f_1^{\mathcal{S}}(u_l))(f_2^{\mathcal{R}}(u_l) + f_2^{\mathcal{S}}(u_l)) \end{aligned}$$

Parameters: Two parties, the receiver $\mathcal{R}$ and the sender $\mathcal{S}$. A field $\mathbb{F}$ and polynomial degrees m, n.
Inputs:

- $\mathcal{R}$: $f_1^{\mathcal{R}}(X) \in \mathbb{F}[X]^{\leq m}$ and $f_2^{\mathcal{R}}(X) \in \mathbb{F}[X]^{\leq n}$.
- $\mathcal{S}$: $f_1^{\mathcal{S}}(X) \in \mathbb{F}[X]^{\leq m}$ and $f_2^{\mathcal{S}}(X) \in \mathbb{F}[X]^{\leq n}$.

Procedure:

(1) $\mathcal{R}$ and $\mathcal{S}$ agree on $L := m+n+1$ distinct evaluation points $\{u_l\}_{l=1}^{m+n+1} \subset \mathbb{F}$.
(2) For $i \in \{1,2\}$ and $\mathcal{P} \in \{\mathcal{R}, \mathcal{S}\}$, $\mathcal{P}$ computes $\mathsf{MEval}(f_i^{\mathcal{P}}, \{u_l\}_{l=1}^{L}) \to (z_{i,l}^{\mathcal{P}})_{l=1}^{L}$.
(3) For $l \in [L]$, $\mathcal{R}$ and $\mathcal{S}$ invoke $\mathcal{F}_{\mathsf{OLE}}$ two times on $z_{1,l}^{\mathcal{R}}, z_{2,l}^{\mathcal{S}}$, and $z_{2,l}^{\mathcal{R}}, z_{1,l}^{\mathcal{S}}$. After, $\mathcal{R}$ and $\mathcal{S}$ obtain $(w_{1,l}^{\mathcal{R}}, w_{2,l}^{\mathcal{R}})$ and $(w_{1,l}^{\mathcal{S}}, w_{2,l}^{\mathcal{S}})$, respectively.
(4) $\mathcal{P} \in \{\mathcal{R}, \mathcal{S}\}$ computes $w_l^{\mathcal{P}} \leftarrow z_{1,l}^{\mathcal{P}} z_{2,l}^{\mathcal{P}} + w_{1,l}^{\mathcal{P}} + w_{2,l}^{\mathcal{P}}$ for $l \in [L]$, $\mathsf{Interp}(\{u_l\}_{l=1}^{L}, \{w_l^{\mathcal{P}}\}_{l=1}^{L}) \to f^{\mathcal{P}}(X)$. Finally, $\mathcal{P}$ outputs $f^{\mathcal{P}}(X) \in \mathbb{F}[X]^{\leq m+n}$.

Figure 14: Protocol $\Pi_{\mathsf{SSrOLE}}^{m,n}$ from $\mathcal{F}_{\mathsf{OLE}}$-hybrid model.

Thus, $f^{\mathcal{R}}(X)+f^{\mathcal{S}}(X)=(f_1^{\mathcal{R}}(X)+f_1^{\mathcal{S}}(X))(f_2^{\mathcal{R}}(X)+f_2^{\mathcal{S}}(X))$. In addition, $w_{1,l}^{\mathcal{P}}, w_{2,l}^{\mathcal{P}}$ follow the uniform distribution over $\mathbb{F}$, so does $w_l^{\mathcal{P}}$ for $\mathcal{P}\in\{\mathcal{R},\mathcal{S}\}$. Hence, $\{w_l^{\mathcal{P}}\}_{l=1}^{m+n+1}$ follows the uniform distribution in $\mathbb{F}^{m+n+1}$, and its interpolation $f^{\mathcal{P}}(X)$ is uniformly distributed over $\mathbb{F}[X]^{\leq m+n}$.

Appendix B. Omitted Proofs

B.1. Proof of Theorem 1

Proof. For $j \in [d]$ and $v \in \mathcal{U}$, define the random variable $\mathfrak{U}_{j,v}$ as the number of elements $x_i = (x_{i,j})_{j=1}^{d} \in X$, $i \in [M]$, s.t. $x_{i,j} \in [v-\delta, v+\delta] \cap \mathbb{Z}$. $\mathfrak{U}_{j,v}$ follows the binomial distribution $\mathrm{Bin}(M, p_{\mathcal{D}_j}(v;\delta))$. If we denote E_j as an event that the number of overlaps at j^{th} axis exceeds C, we obtain

$$\Pr[E_j] = \Pr[\max_{v\in\mathcal{U}} \mathfrak{U}_{j,v} > C] \leq \sum_{v\in\mathcal{U}} \Pr[\mathfrak{U}_{j,v} > C].$$

By applying the Chernoff bound (Lemma 1) on $\mathfrak{U}_v$, we obtain $\Pr[\mathfrak{U}_{j,v} > C] \leq (\frac{eMp_{\mathcal{D}_j}(v;\delta)}{C})^C$. Therefore, we have

$$\Pr[\bigvee_{j=1}^{d} E_j] \leq \sum_{j=1}^{d} \Pr[E_j] \leq \sum_{j=1}^{d}\sum_{v\in\mathcal{U}} (\frac{eMp_{\mathcal{D}_j}(v;\delta)}{C})^C.$$

Since $\Pr[\bigwedge_{j=1}^{d}(\neg E_j)]=1-\Pr[\bigvee_{j=1}^{d} E_j]$, we are done. □

B.2. Proof of Corollary 1

Proof. By plugging the local density $p_{\mathcal{D}_{\mathrm{Unif}}}(v;\delta) = \frac{2\delta+1}{2^u}$ to Theorem 1, we obtain for $\mu = \frac{M(2\delta+1)}{2^u}$,

$$\Pr[X \text{ is } C\text{-max collision}] \geq 1 - d\cdot 2^u \cdot (\frac{e\mu}{C})^C.$$

Here, we can derive that $d\cdot 2^u\cdot(\frac{e\mu}{C})^C \leq 2^{-\kappa}$ if and only if

$$\kappa + u + \log d \leq C(\log C - \log\mu - \log e).$$

If we substitute $\omega = \ln C$, then we obtain $\log C = \frac{\ln C}{\ln 2}$ and

$$C(\log C - \log\mu - \log e) = \frac{e\mu}{\ln 2} e^{\omega - \ln e\mu}(\omega - \ln e\mu).$$

Hence, if we denote $A = \kappa + u + \log d$, we obtain the inequality $e^{\omega-\ln e\mu}(\omega - \ln e\mu) \geq \frac{A\ln 2}{e\mu}$, and for the Lambert W function [38] of branch $x \geq -e^{-1}$ and $\tilde{A} := \frac{A\ln 2}{e\mu}$,

$$C \geq e^{W(\tilde{A})+\ln e\mu} = \frac{e\mu}{W(\tilde{A})}\cdot W(\tilde{A})e^{W(\tilde{A})} = \frac{A\ln 2}{W(\tilde{A})}.$$

Since $W(x) = \Theta(\log x - \log\log x)$ [38] and $A = \Theta(\log d)$, $C = O(\frac{\log d}{\log\log d})$ suffices with other parameters fixed. □

B.3. Proof of Lemma 2

Proof. For each $v \in \mathcal{U}$, let us denote $m(v)$ and $M(v)$ be the smallest and largest indices such that $z_i \in [v-\delta, v+\delta]$. Since each z_i is already sorted, we observe that

$$M(v) - m(v) + 1 = |\{i \in [M] \mid z_i \in [v-\delta, v+\delta]\}|$$

Note that $|z_{M(v)} - z_{m(v)}| \leq 2\delta$. Meanwhile, for indices $i^* > j^*$ that maximizes $i^* - j^*$ under the condition that $|z_{i^*} - z_{j^*}| \leq 2\delta$, we can select the corresponding v by taking the middle point of z_{i^*} and z_{j^*}. Hence, maximizing $M(v) - m(v)$ subject to v is equivalent to direct maximization on the difference between indices. This completes the proof. □

B.4. Proof of Theorem 2

Proof. Observe that $\mathcal{S}$ does not obtain any outputs from subroutines $\mathcal{F}_{\mathsf{OSSI}}$ and $\mathcal{F}_{\mathsf{OT}}$, and the selected random permutation $\pi : [N] \to [N]$ is input-independent. In addition, the proposed membership encoding ensures that $\sum_{v\in\mathcal{U}}\sum_{j=1}^{d} |M_j(v)| = (2\delta+1)dM$. Hence, it suffices to show that $\mathcal{R}$'s view can be simulated from $\mathcal{R}$'s input/output of $\mathcal{F}_{\mathsf{FPSI}}$. We focus on simulating $\{S_i\}_{i=1}^{N}$ from the protocol via ideal functionality inputs/outputs $X = \{x_i\}_{i=1}^{M}$ and $\{z_1, \ldots, z_n\}$, respectively. First, sample a random permutation $\psi : [N] \to [N]$, and define for $i \in [N]$ and $k \in [n]$,

$$\tilde{S}_i := \begin{cases} \{m \in [M] : \|x_m - z_k\|_\infty \leq \delta\} & i = \psi(k) \\ \emptyset & \text{otw.} \end{cases}$$

By Lemma 3 and our choice of ψ, the distribution of $\{\tilde{S}_i\}_{i=1}^{N}$ from functionality inputs/outputs is identical to that of $\{S_i\}_{i=1}^{N}$ from the protocol. This completes the proof. □

B.5. Proof of Theorem 3

Proof. We consider two failure modes: (i) $A_i \uplus B_i$ induces multiplicity in $\cap_{i=1}^{d} A_i$ and (ii) it yields an intersection outside of $\cap_{i=1}^{d} A_i$. For the former, we can derive that such probability is less than $\frac{C^{d+1}}{|\mathbb{F}|^d}$; for each $x \in \cap_{i=1}^{d} A_i$, $\Pr[x \in B_i] \leq \frac{C}{|\mathbb{F}|}$, i.e., $\Pr[\forall i \in [d], x \in B_i] \leq (\frac{C}{|\mathbb{F}|})^d$. Since $|\cap_{i=1}^{d} A_i| \leq C$, we obtain the bound $C\cdot(\frac{C}{|\mathbb{F}|})^d = \frac{C^{d+1}}{|\mathbb{F}|^d}$.

For (ii), let $\Delta := \bigcap_{i=1}^{d}(A_i \cup B_i) - \bigcap_{i=1}^{d} A_i$. For each $x \in \mathbb{F}$, define $j(x) = \min\{i \mid x \notin A_i\}$; $j(x) = \infty$ if $x \in \bigcap_{i=1}^{d} A_i$. If we denote $X_j := \{x \in \mathbb{F} \mid j(x) = j\}$, we can partition $\mathbb{F} - \bigcap_{i=1}^{d} A_i$ via $X_1, \ldots, X_d$. This implies that $\Pr[\Delta \neq \emptyset] \leq \sum_{x\in\mathbb{F}} \Pr[x \in \Delta] = \sum_{j=1}^{d} \sum_{x\in X_j} \Pr[x \in \Delta]$. For $x \in X_j$ we observe that,

$$\begin{aligned}\Pr[x \in \Delta] &= \Pr[x \in B_j \wedge x \in \cap_{i\neq j}(A_i \cup B_i)] \\ &= \Pr[x \in B_j]\Pr[x \in \cap_{i\neq j}(A_i \cup B_i)] \\ &\leq \frac{|B_j|}{|\mathbb{F}|}\Pr[x \in \cap_{i\neq j}(A_i \cup B_i)].\end{aligned}$$

If we denote $G_j := \cap_{i\neq j}(A_i \cup B_i)$— it is well-defined since $d \geq 2$—$\sum_{x\in X_j} \Pr[x \in G_j] = \mathbb{E}[|X_j \cap G_j|] \leq |X_j|$. When $j = 1$, we obtain a tighter bound since

$$\sum_{x\in X_1} \Pr[x \in G_1] = \mathbb{E}\left[\left|X_1 \cap \bigcap_{i=2}^{d}(A_i \cup B_i)\right|\right] \leq |A_2 \cup B_2|$$

With $|B_j| = C - |A_j| \leq C$ and $|A_2 \cup B_2| \leq C$, this gives the bound $\Pr[\Delta \neq \emptyset] \leq \frac{1}{|\mathbb{F}|}\left[(C - |A_1|)C + C\sum_{j=2}^{d} |X_j|\right]$. Here, observe that $X_j \subset \cap_{i=1}^{j-1} A_i - \cap_{i=1}^{j} A_i$, as each element of X_j must appear all of $A_1, \ldots, A_{j-1}$. This gives

$$\sum_{j=2}^{d} |X_j| \leq \sum_{j=2}^{d}(|\cap_{i=1}^{j-1} A_i| - |\cap_{i=1}^{j} A_i|) = |A_1| - |\cap_{i=1}^{d} A_i|.$$

Therefore, we finally obtain

$$\begin{aligned}\Pr[\Delta \neq \emptyset] &\leq \frac{1}{|\mathbb{F}|}[(C - |A_1|)C + C(|A_1| - |\cap_{i=1}^{d} A_i|)] \\ &\leq \frac{1}{|\mathbb{F}|}[C(C - |\cap_{i=1}^{d} A_i|)] \leq \frac{C^2}{|\mathbb{F}|}.\end{aligned}$$

Hence, we obtain the claimed bound. □

B.6. Proof of Theorem 4

Proof. All the communication between $\mathcal{R}$ and $\mathcal{S}$ occurs via the ideal functionalities $\mathcal{F}_{\mathsf{soOPPRF}}$ and $\mathcal{F}_{\mathsf{S^3I}}$, and input sizes of $\mathcal{F}_{\mathsf{soOPPRF}}$ for both parties, namely, $V = \sum_{i=1}^{Q} |S_i|$ for $\mathcal{R}$ and $T = Cd$ for $\mathcal{S}$, can be written via protocol parameters. Hence, it suffices to check whether the protocol indeed returns the desired output of $\mathcal{F}_{\mathsf{OSSI}}$ to $\mathcal{R}$. After Π_{OSSI}, $\mathcal{R}$ obtains the intersection of selected sets with dummy values. Since Theorem 3 ensures that such dummy values do not affect the intersection result except for the probability of at most $\frac{C^2}{|\mathbb{F}|} + \frac{C^{d+1}}{|\mathbb{F}|^d}$, with $|\mathbb{F}| \geq 2^\kappa \cdot C^2$, we can conclude that Π_{OSSI} ensures κ-bit statistical security. □

B.7. Proof of Theorem 6

Proof. Observe that $\gcd(f_S \cdot \gamma_1, f_S \cdot \gamma_2) = f_S \cdot \gcd(\gamma_1, \gamma_2)$. That is, analyzing $\gcd(\gamma_1(X), \gamma_2(X))$ suffices. For a monic irreducible polynomial $g(X)$, $g(X) | \gcd(\gamma_1(X), \gamma_2(X))$ implies that $g(X)|\gamma_1(X)$ and $g(X)|\gamma_2(X)$. Hence, we have

$$\Pr[g|\gcd(\gamma_1, \gamma_2)] \leq \Pr[g|\gamma_1 \wedge g|\gamma_2] = \Pr[g|\gamma_1] \cdot \Pr[g|\gamma_2].$$

Here, if we denote d_g as the degree of $g(X)$, then $g(X)|\gamma_i(X)$ holds if and only if $\gamma_i(X) = g(X)\zeta_i(X)$ for some $\zeta_i(X) \in \mathbb{F}[X]^{\leq 2C-|S|-d_g-1}$, $i \in \{1, 2\}$. This implies

$$\Pr[g(X)|\gamma_i(X)] = \frac{|\mathbb{F}|^{2C-|S|-d_g}}{|\mathbb{F}|^{2C-|S|}} = |\mathbb{F}|^{-d_g}.$$

Now we apply the union bound on the degree d_g and all monic irreducible polynomials of degree d_g. Since the number of monic irreducible polynomials of degree d_g is at most $|\mathbb{F}|^{d_g}$, we obtain the following inequality:

$$\Pr[\gcd(\gamma_1, \gamma_2) \neq 1] \leq \sum_{d_g=1}^{2C-|S|-1} \frac{|\mathbb{F}|^{d_g}}{|\mathbb{F}|^{2d_g}} \leq \sum_{d_g=1}^{\infty} \frac{1}{|\mathbb{F}|^{d_g}}.$$

Since R.H.S is at most $\frac{1}{|\mathbb{F}|-1}$, this completes the proof. □

B.8. Proof of Theorem 7

Proof. Since $\mathcal{S}$'s View includes outputs of $\mathcal{F}_{\mathsf{OLE}}$ and input-independent random polynomials, it suffices to simulate $\mathcal{R}$'s view from its inputs/outputs of $\mathcal{F}_{\mathsf{S^3I}}$. Except for $(h_1^{\mathcal{S}}(X), h_2^{\mathcal{S}}(X))$, all its intermediate received messages are outputs of $\mathcal{F}_{\mathsf{OLE}}$. We thus focus on simulating $(h_1^{\mathcal{S}}(X), h_2^{\mathcal{S}}(X))$. If we denote W as the $\mathcal{R}$'s output of $\mathcal{F}_{\mathsf{S^3I}}$, then Theorem 6 implies that the computed $A(X)$ from the protocol is equal to $\prod_{w\in W}(X - w)$ with probability at least $1 - \frac{1}{|\mathbb{F}|-1}$. By Theorem 5, if we sample two independent polynomials $\varphi_1(X), \varphi_2(X) \xleftarrow{\$} \mathbb{F}[X]^{\leq 2C-|W|-1}$, the following two distributions are equivalent:

$$(\Pi_{w\in W}(X - w) \cdot \varphi_1(X), \Pi_{w\in W}(X - w) \cdot \varphi_2(X))$$

$$(h_1^{\mathcal{R}}(X) + h_1^{\mathcal{S}}(X), h_2^{\mathcal{R}}(X) + h_2^{\mathcal{S}}(X))$$

Hence, except for the probability at most $\frac{1}{|\mathbb{F}|-1}$, $\mathcal{R}$ can simulate $(h_1^{\mathcal{S}}(X), h_2^{\mathcal{S}}(X))$ from $\prod_{w\in W}(X-w)\cdot\varphi_i(X) - \tilde{h}_i^{\mathcal{R}}(X)$ for $\tilde{h}_i^{\mathcal{R}}(X) \xleftarrow{\$} \mathbb{F}[X]^{\leq 2C-1}$, $i \in \{1, 2\}$. □

Appendix C. Additional Analyses on C-Max Collision

For a random variable $\mathfrak{X}$ that follows a distribution $\mathcal{D}$,

$$\begin{aligned}\sum_{v\in\mathcal{U}} p_{\mathcal{D}}(v; \delta) &= \sum_{v\in\mathcal{U}} \sum_{x\in\mathcal{U}} \Pr\big[\mathfrak{X} = x \wedge x \in [v - \delta, v + \delta]\big] \\ &\leq \sum_{x\in\mathcal{U}} \Pr[\mathfrak{X} = x] \cdot (2\delta + 1) = 2\delta + 1\end{aligned}$$

Hence, if we denote $p_{\max} := \max_{v\in\mathcal{U}} p_{\mathcal{D}}(v; \delta)$, we obtain

$$\begin{aligned}\sum_{v\in\mathcal{U}} (\frac{eMp_{\mathcal{D}}(v; \delta)}{C})^C &\leq p_{\max}^{C-1}(\frac{eM}{C})^C \cdot \sum_{v\in\mathcal{U}} p_{\mathcal{D}}(v; \delta) \\ &\leq \frac{2\delta + 1}{p_{\max}}(\frac{eMp_{\max}}{C})^C.\end{aligned}$$

Algorithm 1 Verification of C-Max Collision Property

Require: $X \subset \mathcal{U}^d$ such that $|X| = M$, threshold $\delta \in \mathbb{N}$, and $C \in \mathbb{N}$.
Ensure: `TRUE` or `FALSE` for whether C-max collision property holds.
1: Initialize `cnt` $\leftarrow 0$.
2: **for** j from 1 to d **do**
3: Sort and enumerate the jth component of items X, say $(x_{1,j}, \ldots, x_{M,j}) \in \mathcal{U}^M$ s.t. $x_{1,j} \leq \cdots \leq x_{M,j}$.
4: Initialize $s \leftarrow 1$.
5: **for** r from 1 to M **do**
6: **while** $x_{r,j} - x_{s,j} > 2\delta$ **do**
7: Update $s \leftarrow s + 1$.
8: Update `cnt` $\leftarrow \max\{$`cnt`$, r - s + 1\}$
9: If `cnt` $> C$: **return** `FALSE`.
10: **return** `TRUE`

Hence, for $X = \{x_1, \ldots, x_M\}$ sampled from the distribution $\mathcal{D}$ over $\mathcal{U}^d$ with axis-wise marginals $\mathcal{D}_1, \ldots, \mathcal{D}_d$, Theorem 1 implies that for $p_{\max} := \max_{j \in [d], v \in \mathcal{U}} p_{\mathcal{D}_j}(v; \delta)$,

$$\Pr[X \text{ is } C\text{-max collision}] \geq 1 - d\frac{2\delta+1}{p_{\max}}(\frac{eMp_{\max}}{C})^C.$$

Using the technique in Corollary 1, we obtain the following statement. The proof is exactly the same as that of Corollary 1, except for setting $A = \kappa + \log d + \log \frac{2\delta+1}{p_{\max}}$.

Theorem 9. *Let $\mathcal{D}$ be a distribution on $\mathcal{U}^d$ with marginals $\mathcal{D}_1, \ldots, \mathcal{D}_d$, and $p_{\max} := \max_{j \in [d],\, v \in \mathcal{U}} p_{\mathcal{D}_j}(v; \delta)$. For $X = \{x_1, \ldots, x_M\}$ i.i.d. sampled from $\mathcal{D}$, $\mu := Mp_{\max}$, if $A := \kappa + \log d + \log \frac{2\delta+1}{p_{\max}}$ s.t. $A > \frac{e^2\mu}{\ln 2}$, it suffices to choose $C \geq \max\left\{\mu e,\ \frac{A \ln 2}{W(A\mu^{-1}e^{-1}\ln 2)}\right\}$ so that X satisfies the C-max collision property with probability at least $1 - 2^{-\kappa}$.*

Note that this recovers the result of Corollary 1 for the uniform distribution, where $p_{\max} = \frac{2\delta+1}{2^u}$.

Appendix D. Local Verification of C-Max Collision

Algorithm 1 shows how to verify the C-max collision property of a set $X \subset \mathcal{U}^d$ from Lemma 2 in time $O(udM)$. In `lines 6-7`, the algorithm scans the sorted items per component via r, s to find the maximum difference of indices whose components are close within 2δ. The found $r - s + 1$ is then stored in `cnt` on `line 8`, recording the maximum over the index $j \in [d]$ with respect to each axis.

Complexity Analysis. The algorithm consists of two parts that are shared along the d axes: (i) sorting each component, and (ii) scanning r and s over the range from 1 to M. For the former, we can employ radix sort, which requires $O(uM)$ time complexity. For the latter, we can observe that r and s are scanned over $[M]$ and are monotonically increasing. In addition, the number of comparisons for checking $x_{r,i} - x_{s,i} \overset{?}{>} 2\delta$ and $C \overset{?}{>} r - s + 1$ are at most M each, hence at most $2M$ comparisons suffice. Therefore, aggregating over all d axes, Algorithm 1 takes $O(udM)$ comparisons.

Appendix E. Additional Details on Experimental Analyses

E.1. Micro-Optimizations

Fused Multipoint Evaluation and Interpolation. During $\Pi_{\mathsf{S^3I}}$, we invoke $\mathcal{F}_{\mathsf{SSrOLE}}$ several times while traversing the subproduct tree. This involves a sequential repetition of multipoint evaluation (MEval) and interpolation (Interp). Here, if we represent MEval and Interp as the multiplication of the Vandermonde matrix and its inverse, we can fuse Interp and the subsequent MEval in a single step by multiplying the corresponding Vandermonde matrices in advance.

Truncated PRF. The original alternating moduli PRF by Alamati et al. [22] outputs a 128-bit string by default. However, in our protocol across all parameter settings we experimented with, 64-bit outputs suffice to ensure 40-bit statistical security. In addition, the PRF inputs of each party are structured, i.e., $v \| l$ for some $v \in \mathcal{U}$ and $l \in [C]$. Hence, for a PRF F that outputs a 128-bit string, we use its truncated version F_{T} with redundancy; for a PRF key k_{PRF},

$$F_{\mathsf{T}}(k_{\mathsf{PRF}}, v\|l) := \begin{cases} F(k_{\mathsf{PRF}}, v\|j)_{\mathsf{lo}} & \text{if } l = 2j-1 \\ F(k_{\mathsf{PRF}}, v\|j)_{\mathsf{hi}} & \text{if } l = 2j \end{cases},$$

where x_{lo} and x_{hi} denote the low and high 64-bits of the 128-bit string x, respectively. F_{T} is still a secure PRF, as the low 64 bits and high 64 bits of the secure PRF F's output are computationally indistinguishable from independent uniform random strings. Adopting F_{T} reduces the number of (O)PRF evaluations of F by half.

Constant-Time Branchless GCD Computation. We vectorize the GCD computation in $\Pi_{\mathsf{S^3I}}$ via the Bernstein-Yang algorithm [39], which is constant-time and branchless. Since its runtime is data-independent, one can convert it to a shared-input version for extensions discussed in Section 8.

E.2. Offline Cost Analysis

The offline cost involves (i) OKVS encoding by $\mathcal{R}$ and (ii) OT/OLE correlation generation for so-OPPRF and $\Pi_{\mathsf{S^3I}}$. For the OKVS encoding, $\mathcal{R}$ handles $(2\delta+1)dM$ key-value pairs, and the values are masked by the PRF; in RB-OKVS, encoding takes $O(nw + n\log n)$ operations [30] for n key-value pairs and the band window size w. We select w to ensure the encoding failure probability is less than 2^{-40} with the rate 0.91 (i.e., $\epsilon = 0.1$ in their paper). For the OT/OLE correlation, we can break down the number of 1-out-of-2 OTs and OLE over $\mathbb{F}_{2^{64}}$ in $\Pi_{\mathsf{S^3I}}$. The former requires 512 1-out-of-2 OTs in $\mathbb{F}_2$ (or 256 $\mathbb{F}_4$-VOLE) per so-OPRF evaluation [22]. With our truncated PRF optimization, $\frac{dCN}{2}$ so-OPRF evaluations suffice, so $256 \cdot dCN$ 1-out-of-2 OTs are needed. At the final retrieval of $\mathcal{S}$'s items, N 1-out-of-2 OTs that can accommodate $\mathcal{U}^d$, i.e., $u \cdot d$-bit strings is required. Meanwhile, for the OLE correlations in $\mathbb{F}_{2^{64}}$, we need $dN(2C\log C - 2C + 2) + 4dNC$, where the former term

TABLE 4: Offline time breakdown. Each cell reports wall-clock seconds with its share of the total (%).

Subroutines	C=4	C=8	C=16
OLE Gen.	23.18 (17.3)	105.19 (37.8)	209.36 (41.0)
OT Gen.	65.54 (48.9)	131.27 (47.2)	260.62 (51.1)
PRF Mask	20.96 (15.6)	20.95 (7.5)	21.30 (4.2)
OKVS Enc.	14.12 (10.5)	14.20 (5.1)	14.15 (2.8)
Memb. enc.	8.07 (6.0)	4.79 (1.7)	2.85 (0.6)
other	2.28 (1.7)	1.87 (0.7)	1.94 (0.4)
Offline time (s)	134.14	278.28	510.23

accounts for the subproduct tree evaluation and the latter term corresponds to the two random linear combinations.

We also report the measured offline latency in our implementation, with fixed M=N=2^{16}, d=16, and δ=30, and $C = 4, 8, 16$. Tab. 4 reports the result. We can observe that the OT/OLE generation time dominates the overall offline costs, while the time for OKVS encoding and PRF masking in so-OPPRF remains unchanged because it is independent of C. Notably, the membership encoding time reduces as C increases. This is because the number of inspected universe elements per axis diminishes as we enforce the collision.

End-to-End Protocol Cost Comparison. Based on the offline cost analysis, we provide the end-to-end cost comparison with a reference protocol by Yang et al. [16] in Section 7.3 under the parameters we evaluated above. We excluded Piske et al.'s protocol [14] as it requires only a few hundred OTs in the offline phase regardless of the parameter setting; that is, its end-to-end cost is almost the same as the online cost reported in Tab. 3. In addition, it does not scale to $d = 16$ due to 2^d overhead. The results are reported in Tab. 5. Unlike the online cost, our protocol shows worse end-to-end latency. This is largely attributed to the preprocessing cost from OT/OLE; nevertheless, it can be processed in advance while idle. Reducing the offline cost of our protocol would be another interesting future work.

TABLE 5: Comparison on end-to-end computation latency (s) and communication cost (MB).

Prots.	Offline		Online		E2E	
	Comm.	Comp.	Comm.	Comp.	Comm.	Comp.
[16]	393	55.33	5087	38.81	5480	94.14
Ours (C=4)	62	134.14	1542	4.25	1604	138.39
Ours (C=8)	124	278.28	2759	8.42	2883	286.70
Ours (C=16)	247	510.23	5460	18.54	5707	528.77

E.3. Online Cost Breakdown

We focus on the online computation involved in Π_{OSSI}. With fixed M=N=2^{16}, d=16, and δ=30, we conduct a detailed cost breakdown for $C = 4, 8, 16$. We evaluate the amortized online latency per $\mathcal{S}$'s item, as our protocol invokes $\mathcal{F}_{\mathsf{OSSI}}$ per item. Tab. 6 provides the result, showing that the so-OPRF evaluation step dominates the overall cost, followed by the $\mathsf{S}^3\mathsf{I}$ step and OKVS decoding during so-OPPRF. Since the number of so-OPPRF evaluations grows in $O(C)$, as opposed to the $O(C \log C)$ overhead in $\mathsf{S}^3\mathsf{I}$, the portion for $\mathsf{S}^3\mathsf{I}$ increases as C grows.

TABLE 6: Online time breakdown in Π_{OSSI}. Each cell reports wall-clock seconds with its share of the total (%).

Subroutines	C=4	C=8	C=16
so-OPRF	1.30 (43.2)	2.73 (44.0)	5.63 (39.7)
OKVS Dec.	0.79 (26.0)	1.57 (25.2)	3.16 (22.3)
$\mathsf{S}^3\mathsf{I}$	0.93 (30.8)	1.91 (30.8)	5.38 (38.0)
Online total (s)	3.02	6.21	14.17
Per item (μs)	46.07	94.79	216.29

TABLE 7: Online time breakdown in $\Pi_{\mathsf{S^3I}}$. Each cell reports wall-clock seconds with its share of the total (%). Sub-timers are profiling counters; "other" is unattributed ones.

Subroutines	C=4	C=8	C=16
Subproduct-tree	0.55 (58.7)	1.31 (68.6)	4.09 (76.0)
Rand. lin. comb.	0.25 (26.7)	0.46 (24.2)	0.99 (18.4)
GCD	0.05 (5.3)	0.03 (1.5)	0.09 (1.8)
Other	0.09 (9.3)	0.11 (5.6)	0.21 (3.9)
$\mathsf{S}^3\mathsf{I}$ online (s)	0.93	1.91	5.38

We also provide a detailed breakdown of the inside of the $\mathsf{S}^3\mathsf{I}$ in Tab. 7. The subproduct tree evaluation accounts for most of the computation, as it incurs $O(dCN \log C)$ overhead. Meanwhile, the random linear combination part takes $O(dCN)$ overhead, and the same tendency is also observed. The number of GCD computations is $O(N)$, and thanks to the adoption of the Bernstein-Yang algorithm, it takes only a small portion of the whole computation.

Appendix F. Details on Real-World Datasets

NZ-COVID Dataset. The data were collected from the COVID contact location reports by the Ministry of Health, New Zealand. Among them, we used location data spanning 7 months (2021-08-01 – 2022-02-22) from the 708 commits (from `a395fbb` to `eb4d2db`) to `august-2021/locations-of-interest.csv` in their official GitHub repository, obtaining 700 valid snapshots. After removing duplicates in event IDs and rounding-off errors $\leq 10^{-9}$, we finally obtain 2,167 locations; for an ID that appears in multiple snapshots, we use its entry from the most recent snapshot. We then convert each geocoded location to meters via equirectangular projection [40]. We choose the mean latitude of the 2,167 locations as the standard parallel, and assume Earth's radius is 6,371 km.

LANL Dataset. This dataset contains de-identified multi-source cybersecurity events collected by LANL over 58 consecutive days from 12,425 users. Among them, we collect successful interactive authentication events of user accounts (`source user@domain`), which are included in `auth.txt.gz`, and, for each observed day, record the time of day of its first such authentication. We interpret each timestamp by considering the quotient and remainder when divided by 86,400. We select accounts observed on at least 24 days, and regard the corresponding time of day values from the first 24 observed days as a single data item for each account. After that, we finally obtain 1,026 accounts.

Appendix G.
Ethics Consideration

We attest that we thoroughly read the ethics discussions in the conference call for papers, the detailed submission instructions, and the ethics guidelines. We attest that the research team considered the ethics of this research, that the authors believe the research was done ethically, and that the team's next step plans are ethical. All the experiments for efficiency evaluation were conducted with synthetic data. For real-world datasets, we employed publicly available datasets only and clearly documented how they were collected, processed, and utilized in this research.

Appendix H.
Generative AI Usage Considerations

The authors used ChatGPT 5.5 and Claude Opus 4.8 for grammar check, language polish, and cleaning/tidying up the codebase. We attest that all the intellectual contributions—C-max collision property, membership tag-based abstraction, protocol design and instantiation, and proofs—originate from the authors.

Appendix I.
Meta-Review

The following meta-review was prepared by the program committee for the 2027 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

I.1. Summary

This paper addresses a long-standing tension in Fuzzy Private Set Intersection (FPSI): existing symmetric-key protocols either assume strict separation between points (incurring exponential $O(2^d)$ overhead via spatial hashing) or assume disjoint projections across axes (achieving poly(d) overhead, but only statistically justifiable for certain data distributions). The authors introduce a new "density-bounded" structural assumption, C-max collision, which permits points to cluster within a δ-neighborhood as long as the axis-wise overlap multiplicity does not exceed a constant C. They formalize a new cryptographic abstraction, Oblivious Selected Set Intersection (OSSI), and instantiate it using symmetric-key primitives. The resulting protocol targets the ℓ_∞ metric and achieves complexity linear in set size and dimension for fixed C. The paper includes a theoretical analysis of when C-max collision holds, a local verification algorithm for computing C on a given dataset, and an experimental evaluation against representative separation-based baselines.

I.2. Scientific Contributions

- Provides a Valuable Step Forward in an Established Field
- Creates a New Tool to Enable Future Science (the OSSI functionality)

I.3. Reasons for Acceptance

1) The paper resolves a genuine and previously unaddressed tension in the FPSI literature: it offers a structural assumption that is incomparable to, and more flexible than, existing separation-based assumptions, with a graceful (parameterized by C) degradation rather than an all-or-nothing requirement.
2) The core cryptographic construction is technically sound. Reviewers found no flaws in the reductions, the security proofs (modulo one gap discussed below and resolved during rebuttal), or the use of Chernoff bounds to justify the assumption theoretically.
3) The OSSI functionality is a clean, reusable abstraction that reviewers expect could be adopted in future MPC constructions beyond this paper. In response to the most substantive concern raised across multiple reviews — whether C-max collision is justifiable on real-world data, rather than only synthetic uniform samples — the authors provided new empirical evidence during the interactive rebuttal period, evaluating C on two real datasets (NZ COVID contact-tracing location data and LANL enterprise authentication logs). On both, the disjoint-projection assumption used by prior work fails outright, while C-max collision holds with modest, usable values of C. This was a direct and convincing response to reviewer concerns, and shifted at least two reviewers toward support.
4) The authors also provided a rigorous fix, during rebuttal, to a real gap identified in the security proof of Theorem 4 regarding potential leakage of the OSSI input size Q, and added a head-to-head empirical comparison against prior work (Yang et al.) in the regime where both assumptions hold, clarifying the practical trade-off between the two approaches.