Approximate Private Set Intersection

Steve Lu, CEO

Stealth Software Technologies, Inc.

Presented on 2026-10-27 at WPEC 2026
NIST Workshop on Privacy-Enhancing Cryptography 2026

STEALTH
SOFTWARE TECHNOLOGIES

Approximate PSI

- Joint work with Wutichai Chongchitmate and Rafail Ostrovsky
  - Extends our prior approximate PSI line of work <https://eprint.iacr.org/2024/682>
- Déjà vu
  - I gave a talk on approximate PSI 2 years ago at WPEC
  - What has changed?
- Today I hope to share with you new results and perspectives on approximate PSI

Outline

1. **Overview**
2. **New Lens on PSI**
3. **Our Results**
4. **Conclusion**

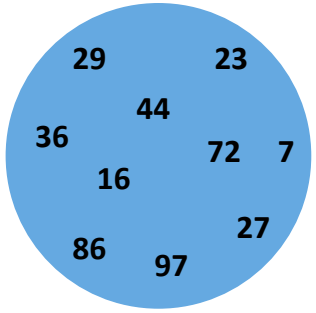
STEALTH
SOFTWARE TECHNOLOGIES

Outline

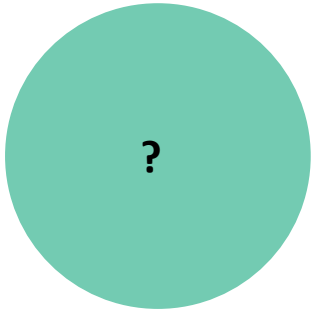
1. **Overview**
2. **New Lens on PSI**
3. **Our Results**
4. **Conclusion**

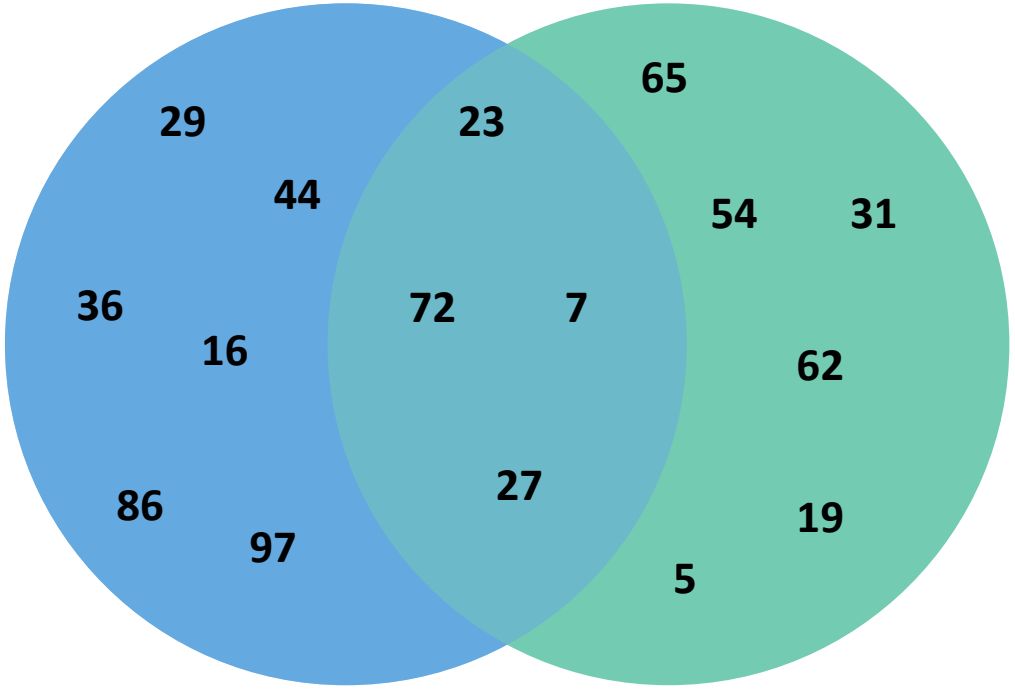
STEALTH
SOFTWARE TECHNOLOGIES

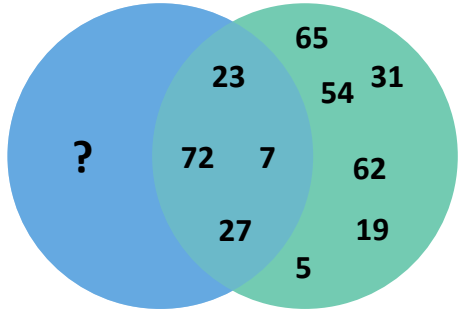
Review: Approximate PSI







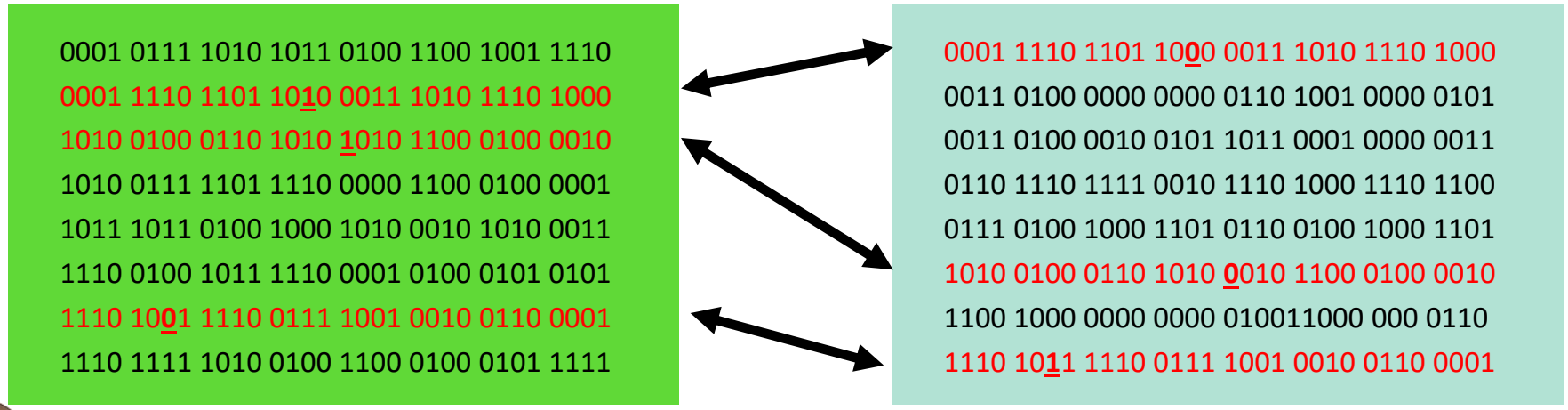






STEALTH
SOFTWARE TECHNOLOGIES

Review: Approximate PSI



- Matching elements when they are “close” with respect to some distance metric
- E.g. Hamming distance between strings

STEALTH
SOFTWARE TECHNOLOGIES

Review: Fuzzy Matching

- Fuzzy matching was proposed even in [FNP04]
- Applications to biometrics drove a lot of these works [CH08, EFGKLT09, SSW09, YSPW09, OPJM10, BG11, HEKM11, UCKBL21]
- “t-out-of-T” or distance matching approach

Review: Structure-Aware PSI [GRS22,GGM24]

- Assuming an input set of one party has certain structure
- Efficiency depends on the structure, not the number of elements
- When the structure is a union of balls (with respect to some distance metric), this implies approximate PSI by replacing the party's input with the balls around each element
- [GRS22]: SA-PSI for ℓ_∞ distance on vectors of u-bit integers

Review: Distance-Aware PSI [CFR23]

- Both parties learn all matched pairs
  - Non trivial to modify to only one party learning their parts of matches
- Constructed through pairwise distance comparison test
  - Resulting in quadratic complexity in the size of input sets
- [CFR23]: DA-PSI for Hamming distance with communication independent of input length (but quadratic in the threshold), with small false negative rate

Additional Works on Fuzzy PSI

- Fuzzy PSI [BP24] where only the receiver learns the output
  - Standard variant (FPSI): the receiver learns the sender's elements that match
  - Sender Privacy variant (FPSI-SP): the receiver only learns their elements that match
- [GQL+25]: FPSI(-SP) for ℓ_p, ℓ_∞ and Hamming distance under unique component assumption
- [BP25]: FPSI(-SP) for ℓ_∞ under no structural assumption but exponential complexity
- [BN25]: FPSI(-SP) for Hamming distance under no structural assumption but quadratic computation for the receiver

Our Contribution

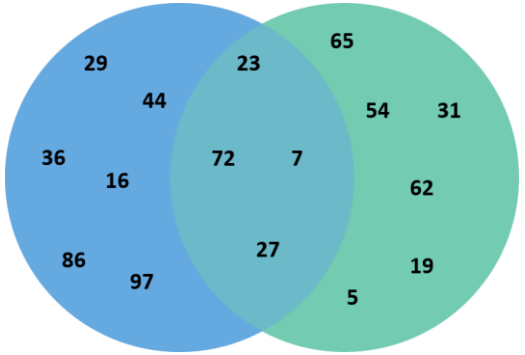
- When our original Approximate PSI came out in 2024, the Hamming distance variant was 20x faster and used 30% less communication in various regimes
- Since then, many works in Fuzzy PSI have emerged
- In this talk, I want to first focus on contextualizing Private Set Intersection, then talk about new results

Outline

1. **Overview**
2. **New Lens on PSI**
3. **Our Results**
4. **Conclusion**

STEALTH
SOFTWARE TECHNOLOGIES

Perspective on Private Set Intersection



Is this how we want to think of PSI?

- Let's look at not just applications but classes of applications of PSI
- Going from exact to approximate enables new PSI opportunities

STEALTH
SOFTWARE TECHNOLOGIES

PSI As Privacy-Preserving Record Linkage

- Hash-based PPRL: Sender and receiver use a seeded hash to hash their data, a third party who doesn't know the seed compares and builds the linkage for them
- Hashes are by-and-large exact match, but combined with a threshold… (for example, in [KCJ+15])

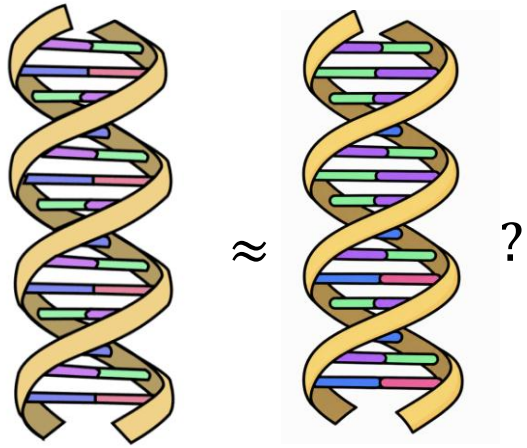
- This is coarse-grained approximate matching!
- PSI can remove the need for a third party

> In this initial implementation, we matched records using the following simple deterministic algorithm. Two records from different sites are considered to "match" (i.e., be associated to the same patient) if they have the same:
>
> 1. Seeded HashID of (First Name + Last Name + Date of Birth),
> 2. Seeded HashID of (Date of Birth + SSN),
> 3. Seeded HashID of (Last Name + SSN), or
> 4. Seeded HashID of (Three Letter First Name + Three Letter Last Name + Soundex First Name + Soundex Last Name + Date of Birth + SSN).
>
> These four matching criteria were selected because
>
> - the necessary data are available in almost all medical records,
> - it is straightforward to understand what constitutes a match, and
> - false positives are very unlikely.

STEALTH
SOFTWARE TECHNOLOGIES

PSI As Similarity Matching

- Merging geospatial and demographic data
- Healthcare/genomics
  - Detect the same person across biobanks
  - Usually require thousands pairs of SNPs
  - Genotyping error (up to 1%)
  - The difference between two people is large, even for relatives.
- Imaging and biometrics
- Text similarity



PSI As Retrieval and Membership

- Did something big happen in the past few years?
- High-dimensional nearest-neighbor, with privacy
  - Retrieval Augmented Generation
  - Private augmentation of Large Language Models
  - Near-matching a vector to a large database
- Membership testing
  - Traditionally known as unbalanced PSI
  - One side wants to test one or a few things rather than set vs set
  - Contact tracing, patient discovery
  - Our new work introduces a new approximate membership test

Outline

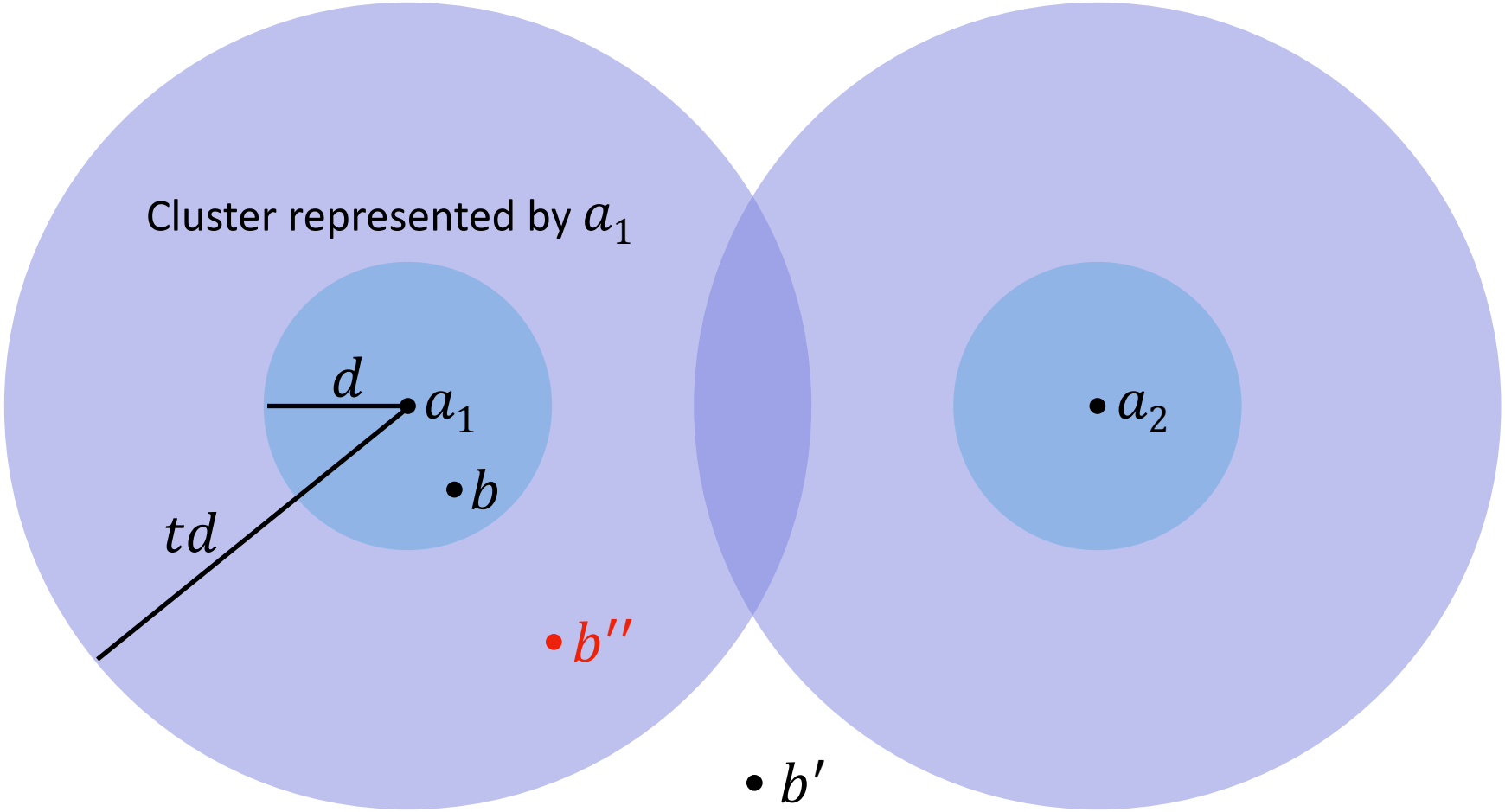
1. **Overview**
2. **New Lens on PSI**
3. **Our Results**
4. **Conclusion**

STEALTH
SOFTWARE TECHNOLOGIES

Review: Assumption for Our Setting [CLO24]

- When the threshold is sufficiently large, all n^2 pairs of elements in the input sets are potentially matched
- In the exact match setting, matches are one-on-one
  - Can we have something in-between?
- We consider a setting where elements are either near (within threshold) or far (multiple of the threshold, i.e. “gap” times threshold)

Review: Near or Far



STEALTH
SOFTWARE TECHNOLOGIES

Review: Approximating Matching [KOR08]

- "Locality preserving hashing" via projection mapping
- For elements of length ℓ , choose a random index set $I \subseteq [\ell]$
  - Choose $i \in [\ell]$ to be in I independently with probability p
- Pairs that are close are more likely to collide than pairs that are far apart

$$I = \{1,3,7,10,16,17,26,29\}$$

a =0001 1110 1101 1010 0011 1010 1110 1000

b = 0001 1110 1101 1000 0011 1010 1110 1000

a' =0001 0111 1010 1011 0100 1100 1001 1110



a_I =0011 0011

$=$

b_I =0011 0011

$\neq$

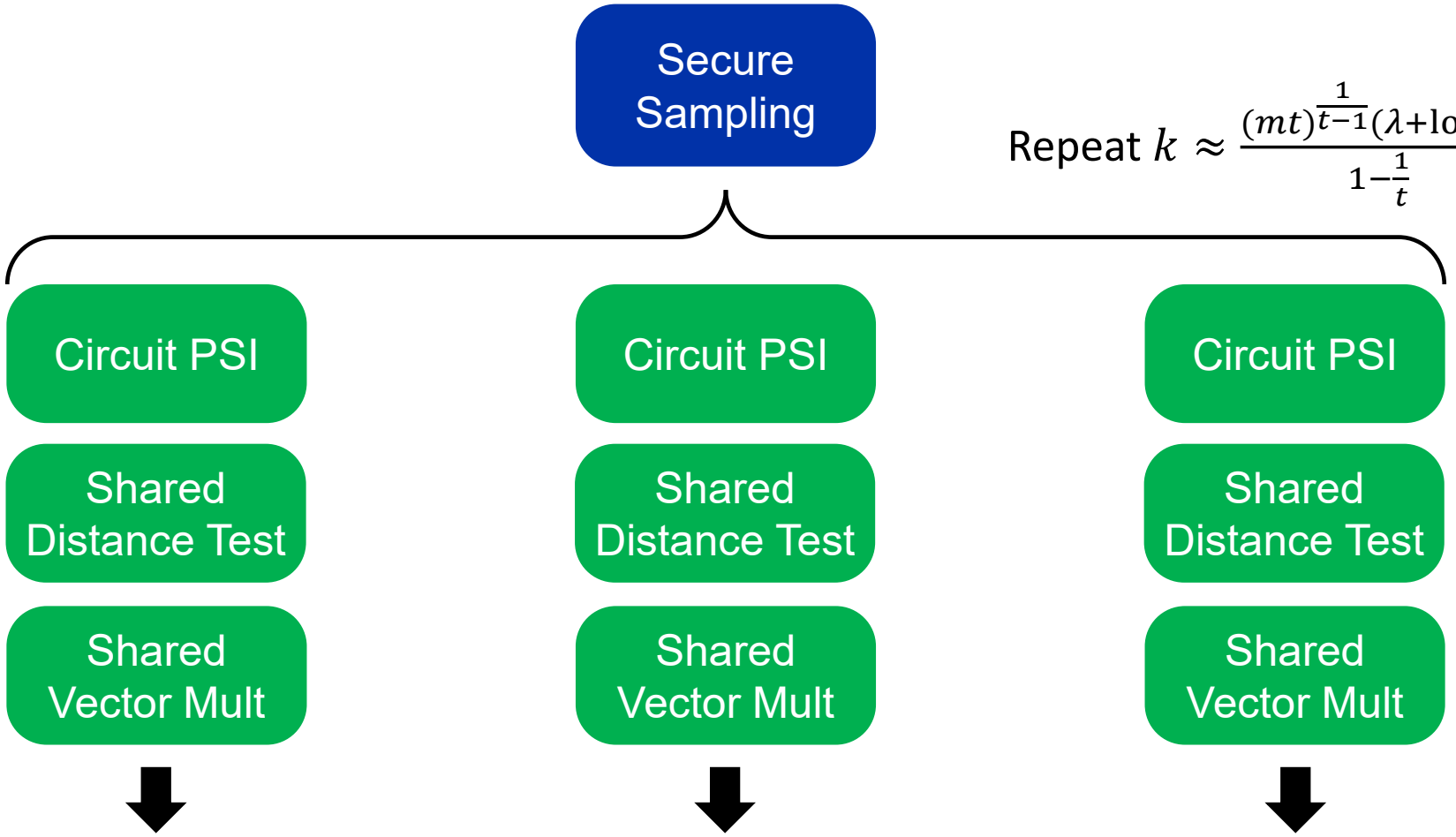
a'_I =0010 1001

Repeat k times, indep.

Review: Overall Flow [CLO24]

- Securely sample projection positions into projected sets
- Compute standard PSI on each (smaller) projected set and output actual elements of the matches as secret shares
    - Circuit PSI
- Distance comparison test that outputs secret shares (Hamming)
- Use the test results as selection vector in each projected set to eliminate non-matching pairs (pointwise multiplication)

Review: Approx-PSI Protocol [CLO24]



STEALTH
SOFTWARE TECHNOLOGIES

New Improvements

- We successfully melded the secure distance testing with the circuit-PSI computation itself
- Adopted recent community improvements to secure vector multiplication protocols
- Introduced a new unbalanced approximate PSI for approximate search/membership testing

New Private Approximate Membership Testing

- Naïve solution: use secure distance test against all elements O(nm)
- Improve using bucketing and private information retrieval (PIR) techniques [KM21]
- Small overhead over the underlying PIR
  - Sublinear in the set size
- Exponentially small False Negative Rate
- No gap or structural assumption needed

New Comparison: Hamming Distance

Term	Meaning
S	Sender
R	Receiver
m	Sender set size
n	Receiver set size
d	Distance
p	Special Norm
ℓ	Length of data (dimension)
t	Assumed gap in non-matching data
B1, B2	Fully Homomorphic Encryption Noise Parameters
λ	Security parameter

Prot.	Assumption	Comm.	S. Comp.	R. Comp
[UZK+21]	FPR&FNR	$\mathcal{O}(B_1\ell mn)$	$\mathcal{O}(B_2\ell mn)$	$\mathcal{O}(\binom{\ell}{d}n)$
[CFR23]	-	$\mathcal{O}(d^2mn)$	$\mathcal{O}((\ell+d^2)mn)$	$\mathcal{O}(\ell mn)$
[BP24] *,†	R far $\geq td$	$\mathcal{O}(d\ell n^{1+\frac{1}{t}} + dmn^{\frac{1}{t}}\log n)$		$\mathcal{O}(\ell mn^{\frac{1}{t}}\log n)$
[BP25] *,†	S&R far $\geq 2d\ell^{\frac{1}{p}}$	$\mathcal{O}(2^\ell n(\ell + p\log d) + m(d\ell + p\log d))$		
[GQL+25] †	R unique	$\mathcal{O}(\ell^2 m + \ell dn)$	$\mathcal{O}(\ell^2 m)$	$\mathcal{O}(\ell^2 m + \ell dn)$
[BN25]	-	$\mathcal{O}(\ell m + \ell dn)$		$\mathcal{O}(\ell dmn)$
Ours, improved from [CLO24]	S/R far $\geq td$	$\mathcal{O}(\ell(m+n)m^{\frac{1}{t-1}}\log n)$		$\mathcal{O}(\ell nm^{\frac{1}{t-1}}\log n)^{\ddagger}$
	$t = \mathcal{O}(\log m)$	$\mathcal{O}(\ell(m+n)\log n)$		$\mathcal{O}(\ell n\log n)^{\ddagger}$

- Our protocol is the only one without factor of ℓ^2, ℓd or d^2 in both communication and computation.

STEALTH
SOFTWARE TECHNOLOGIES

Experimental Results

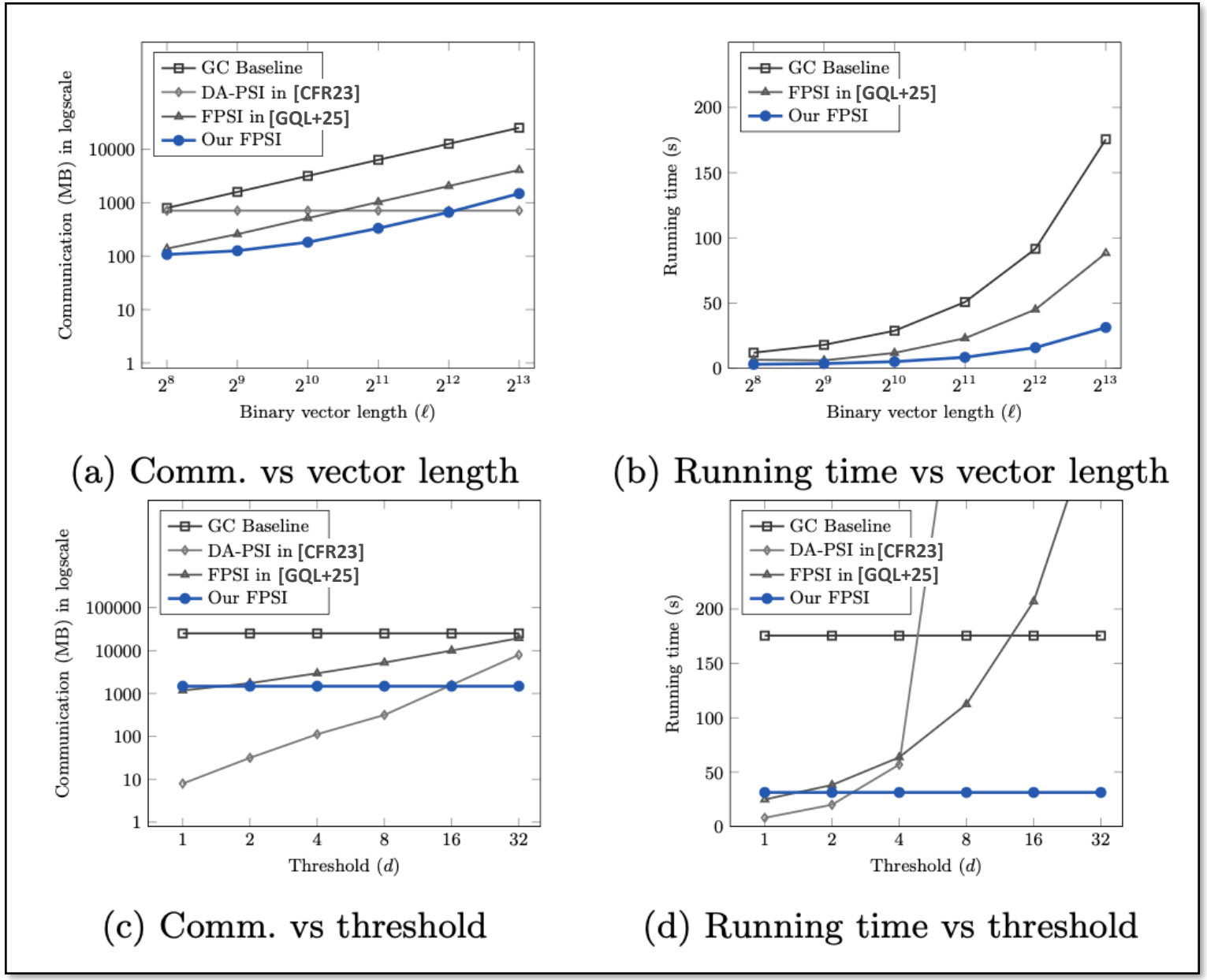
We tested our protocol on randomly generated sets consisting of ℓ = 128-bit vectors of set sizes $n = m = 256, 1024, 4096,$ threshold $d = 4$, gap $t = \log n$, statistical security parameter $\lambda = 40$

Prot./Steps	256		1024		4096	
	Comm.	Time	Comm.	Time	Comm.	Time
Projection	0.01	0.004	0.01	1.44	0.01	4.84
spssPSI						
- ssFMatchT	126.87	3.71	491.9	14	1980	57.2
- ssVMult	4.51	0.648	16.65	2.274	64.35	8.51
spssPSI total	133.09	4.938	514.33	18.154	2066.4	72.81
Open & output	3	0.35	11.05	1.21	41.89	4.62
Our FPSI	136.10	5.292	525.39	20.804	2108.3	82.27
Quad. FPSI	150.63	8.29	2410	133.4	-	-

Time is in seconds

Communication is in MB

STEALTH
SOFTWARE TECHNOLOGIES



- For fixed d=6, our protocol communicates less up to ℓ = 4096 while having the best running time. DA-PSI is not shown in (b) as it is too large.
- For fixed ℓ = 8192, our protocol outperforms others when d≥16.

STEALTH
SOFTWARE TECHNOLOGIES

Outline

1. **Overview**
2. **New Lens on PSI**
3. **Our Results**
4. **Conclusion**

STEALTH
SOFTWARE TECHNOLOGIES

Conclusion & Key Takeaways

- New and improved Approximate PSI
  - Useful for linkage, matching, retrieval
  - Extends to angular distance/Cosine similarity, Euclidean distance and edit distance through transforms
- New Approximate Membership Testing
  - Sublinear communication in the set size
- Post-quantum from post-quantum building blocks