

Paths Toward PSI Standardization and Approximate PSI

Steve Lu, CEO

steve@stealthsoftwareinc.com

Stealth Software Technologies, Inc.

NIST Workshop on Privacy-Enhancing Cryptography 2024
2024-09-24 (PSI Day)



Part I: Approximate PSI

Approximate PSI

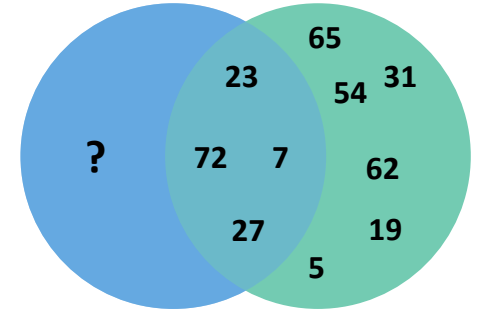
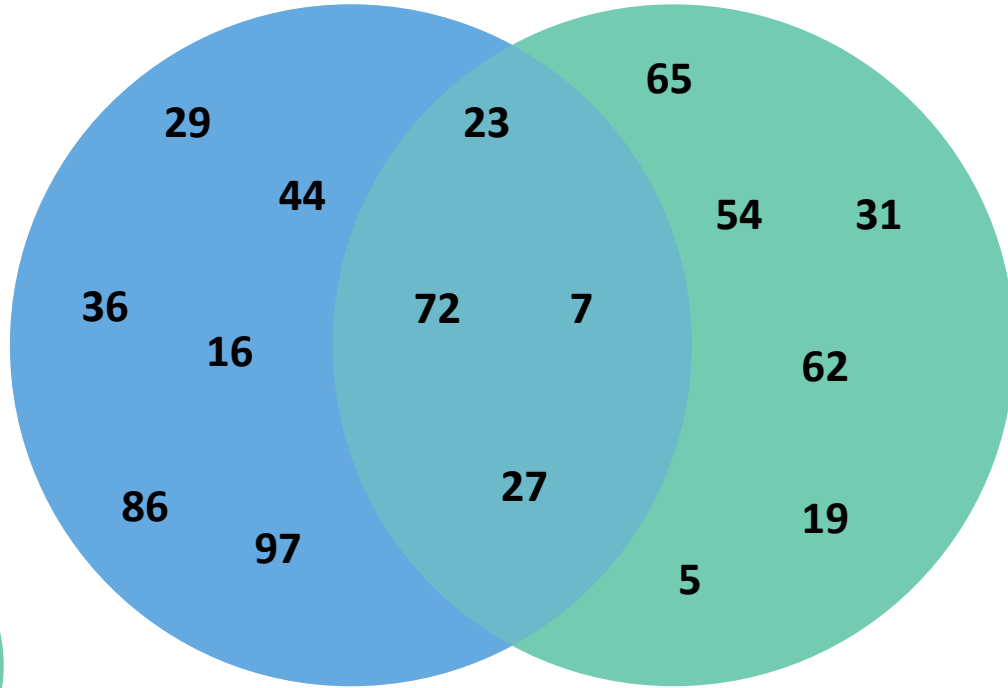
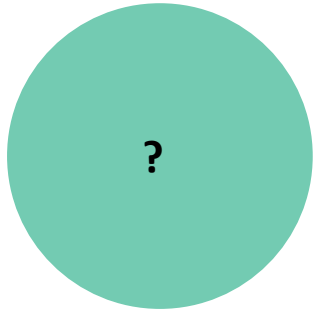
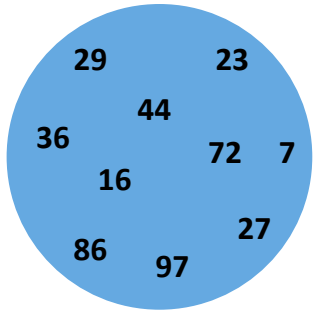
- “Approximate PSI with Near-Linear Communication”
- Joint work with Wutichai Chongchitmate and Rafail Ostrovsky
- <https://eprint.iacr.org/2024/682>

Outline

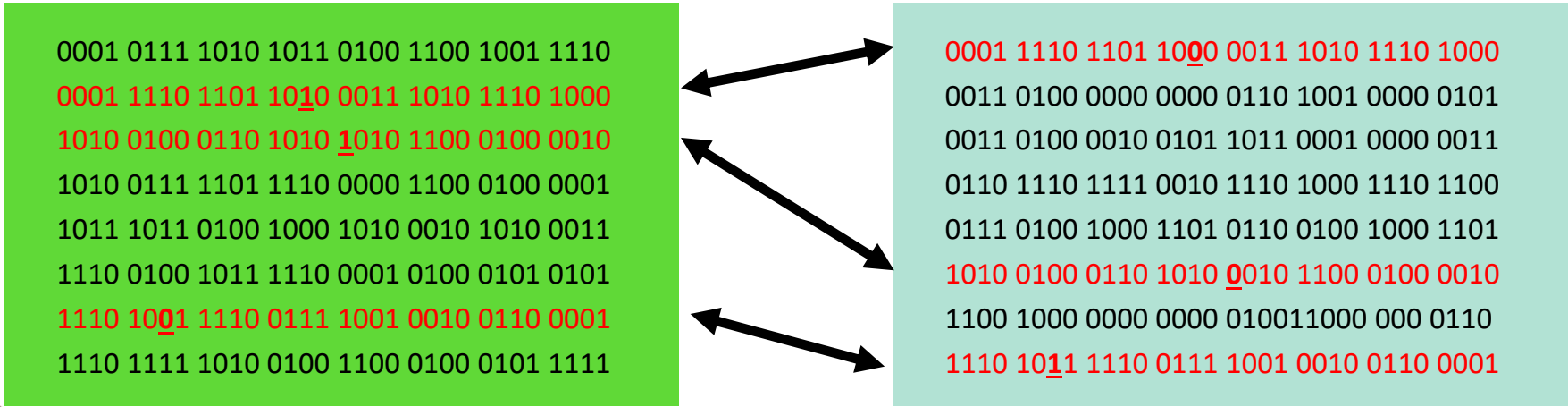
- Overview
- Our Techniques
- Benchmarks
- Remarks

Outline

- **Overview**
- Our Techniques
- Benchmarks
- Remarks



Approximate PSI



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



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

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

Distance-Aware PSI [CFR23]

- Similar setting as ours
- 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)

Our Contribution

- Distance-based PSI, where nearby pairs of elements are returned (to one or both parties) according to some metric
- We **restrict the gap** between elements to avoid too many matches
- With this restriction, we are able to achieve linear (up to polylog factors) communication using circuit PSI as a building block
- When considering Approximate Hamming PSI, ours is **20x faster and 30% less communication** on many different regimes of set sizes and thresholds

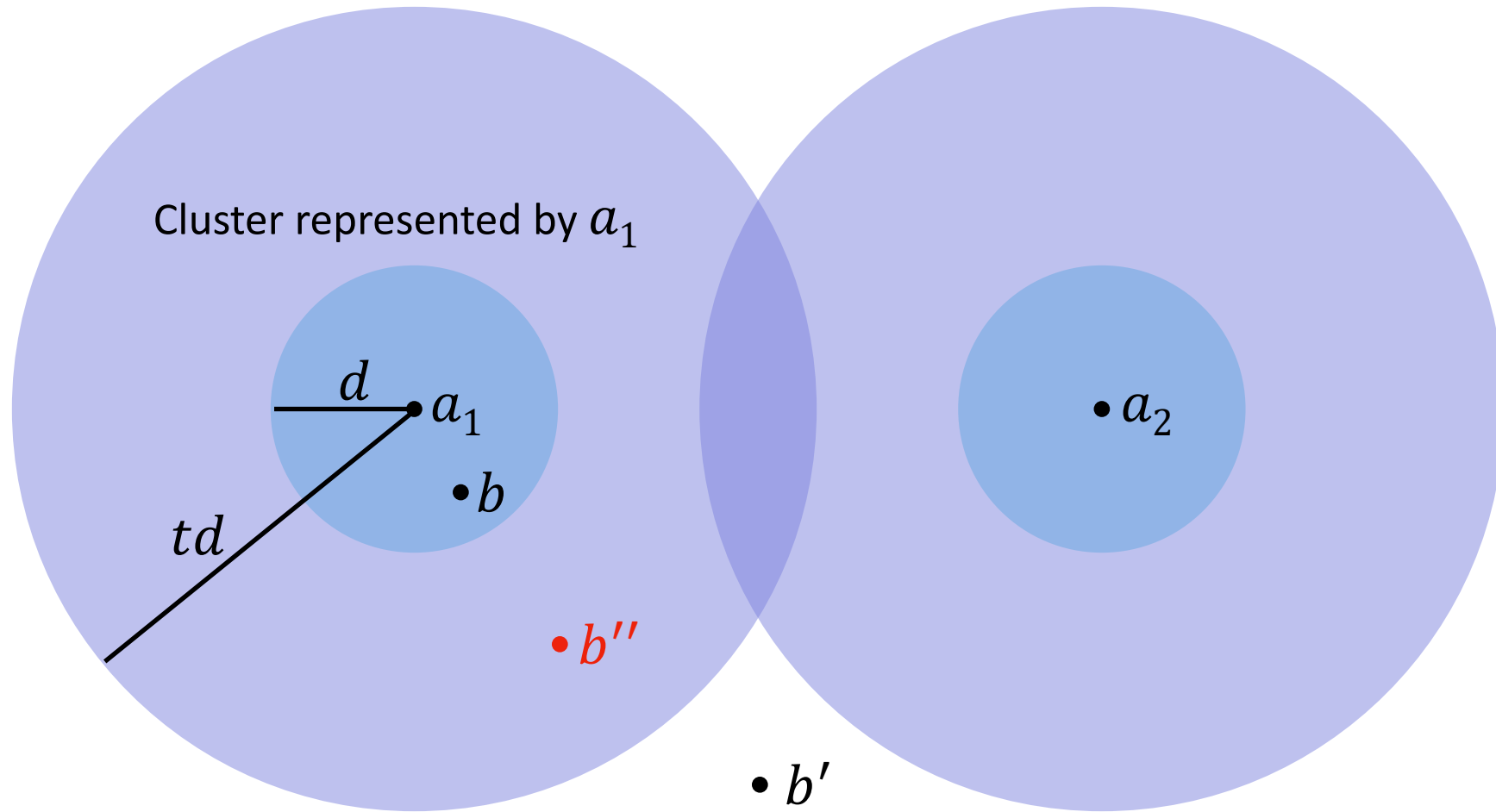
Outline

- Overview
- **Our Techniques**
- Benchmarks
- Remarks

Our Setting: Near or Far

- 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)

Setting: Near or Far



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

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

$a = 0001\ 1110\ 1101\ 10\underline{1}0\ 0011\ 1010\ 1110\ 1000$

$b = 0001\ 1110\ 1101\ 10\underline{0}0\ 0011\ 1010\ 1110\ 1000$

$a' = 0001\ 0111\ 1010\ 1011\ 0100\ 1100\ 1001\ 1110$



$a_I = 0011\ 0011$

$=$

$b_I = 0011\ 0011$

\neq

$a'_I = 0010\ 1001$

Repeat k
times,
indep.

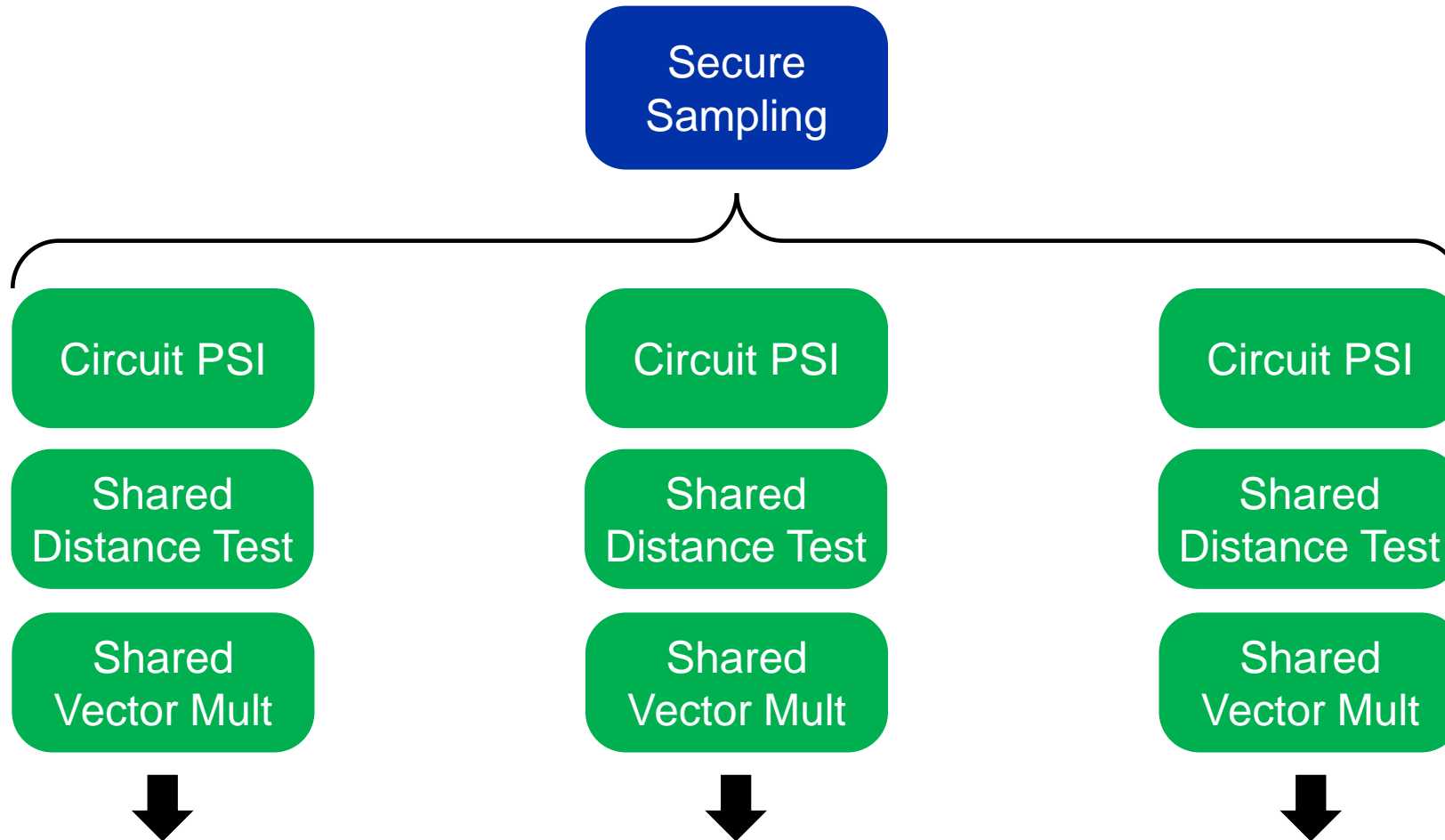
Non-Matching Analysis

- We hope that projections of the non-matching pairs (far apart) do not collide with high probability
- Number of trials needs to be exponential in security parameter/ $\log n$
- Thus, in general, we need to further test for non-matching pairs that collide
 - Good news, this is mostly near linear

Overall Flow

- 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)

Approx-PSI Protocol



Additional Improvements

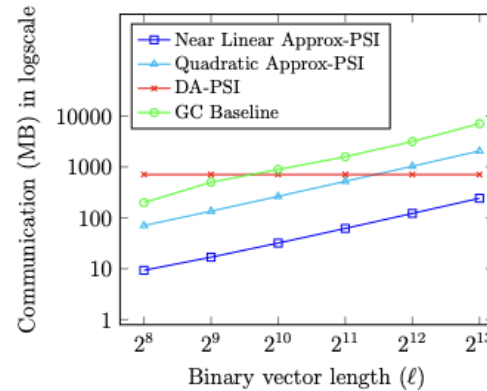
- Merge additional tests into the circuit of the circuit-PSI computation
- New gap-Hamming test protocol
- and more...

Outline

- Overview
- Our Techniques
- **Benchmarks**
- Remarks

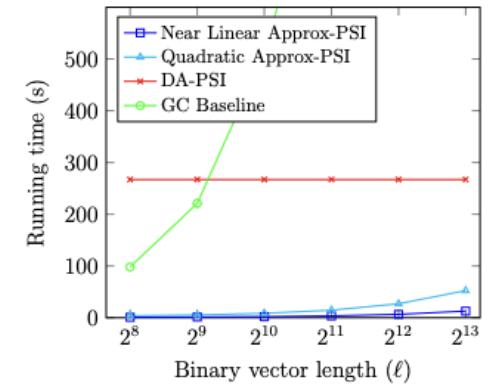
Main Result: Approx-PSI for Hamming distance

- Compared to DA-PSI
 - Adjusted security parameter to match 5% error rate for apples-to-apples comparison
- More efficient in communication when input lengths are not too large, or when the threshold is not too small
- Significantly faster in all settings

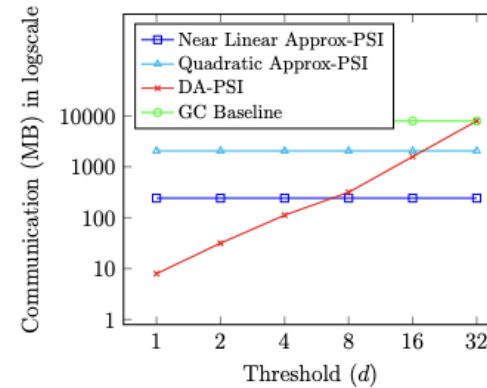


Fixed $d = 6$,
 $t = 16$, $\lambda = 5$

(a) Comm. vs vector length

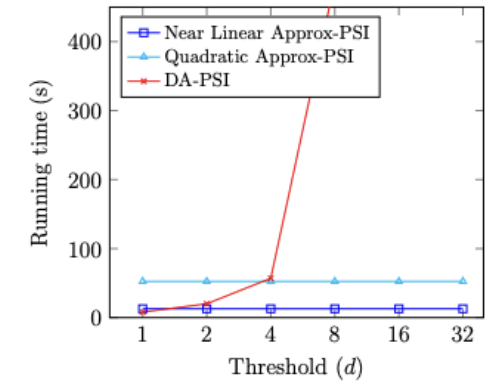


(b) Running time vs vector length



Fixed $\ell = 8192$,
 $t = 16$, $\lambda = 5$

(c) Comm. vs threshold



(d) Running time vs threshold

Main Result: Approx-PSI for Hamming distance

Step	communication (MB)			running time (s)			
	n	256	1024	4096	256	1024	4096
SS Ham PSI		146.13	552.96	2180	10.67	40.04	156.93
SS Vector Mult.		3.867	14.34	55.45	2.693	9.969	37.8
Open & output		0.476	1.765	6.824	0.258	0.955	3.83
Total		150.48	569.07	2242.3	13.62	50.96	198.56
Quadratic Approx-PSI		151.44	2426.75	-	9.558	154.13	-

Fixed $\ell = 128, d = 4, t = \log n, \lambda = 40$

Outline

- Overview
- Our Techniques
- Benchmarks
- Remarks

Approx-PSI for Other Distance Metrics

- Embed real/binary vectors using low distortion embedding from other distance metrics
- The gap setting preserves near and far pairs through the embedding
- The embeddings from Euclidean distance and angular distance based on Johnson-Lindenstrauss lemma, improved through a line of studies [PV14, OR15, HS20, DS20, DM21]
- The embedding from edit distance is from [OR07]

Conclusion

- Approx-PSI for Hamming distance
 - Near linear complexity
 - Improved concrete efficiency, especially in running time
 - New gap-Hamming distance comparison test
- Approx-PSI for angular distance, Euclidean distance and edit distance through transforms

Part II: Standardizing PSI

Tradeoff Dimensions

Structure

2-party
N-party, all-wise
N-party, pair-wise
With trusted party

Approaches

Diffie-Hellman/RSA/OPE
Oblivious PRF
OT/(V)OLE
Circuit
FSS

Adversary Model

Semi-honest
Malicious-with-abort
Malicious-*
Size-revealing

Sizes

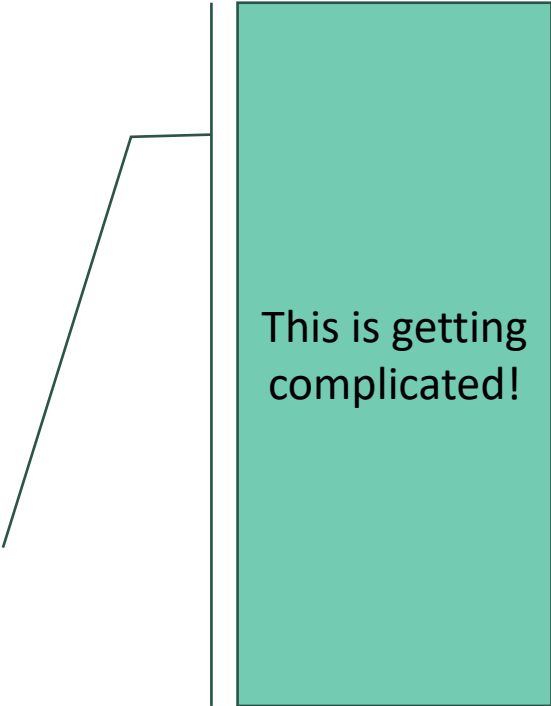
Asymmetric
Symmetric Large
Symmetric Small

Security Assumption

Diffie-Hellman
LPN-like
MPC/FHE

Function

Plain Intersection
Cardinality
Union
Intersection-Sum
Approximate



This is getting complicated!

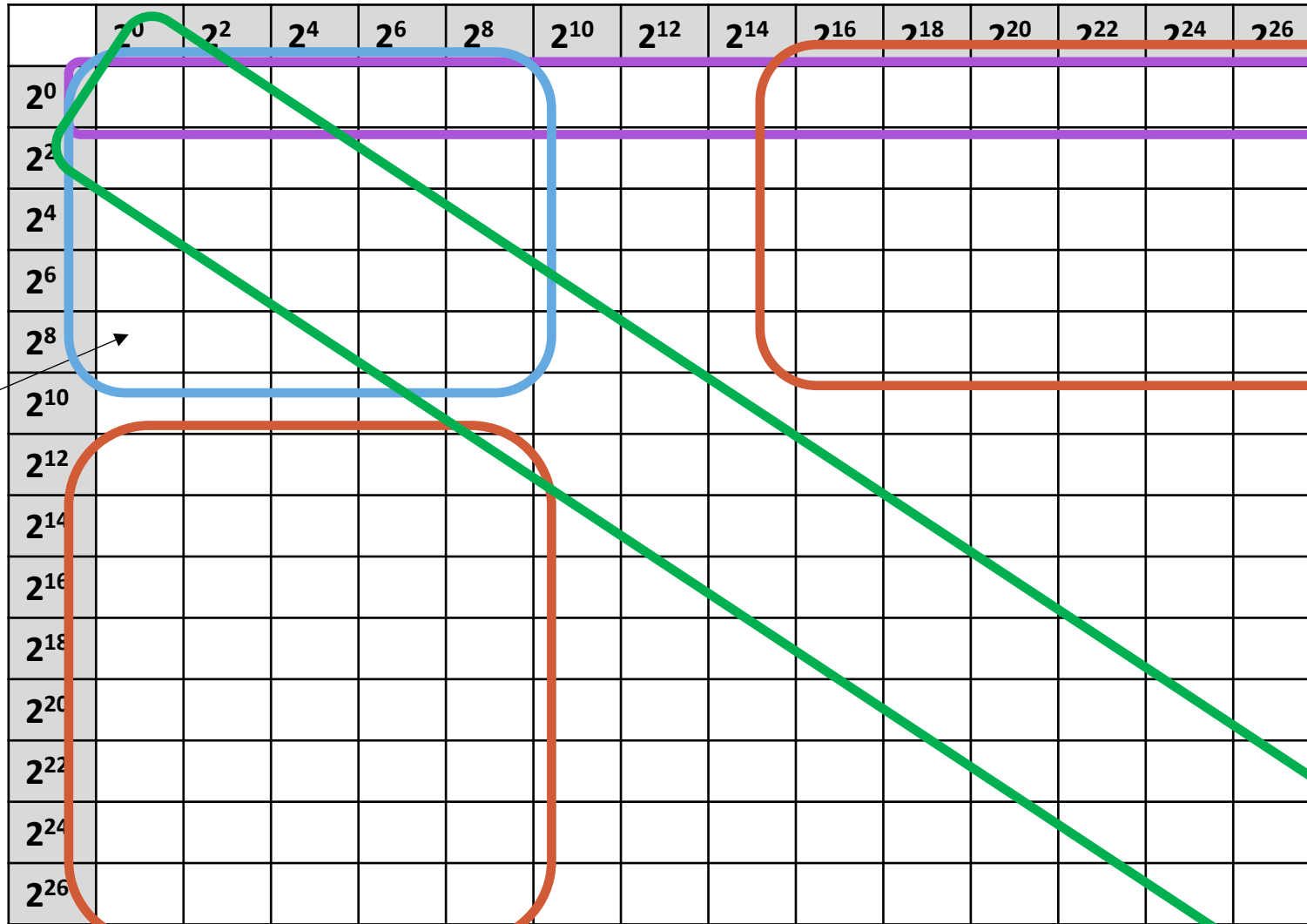
A Modest Proposal

- Personally, I would like to see three undertakings:
 1. Living document on raw performance
 - Pick a simple setting, e.g. Alice & Bob exact intersection, only Alice gets answer
 - Vary set sizes of Alice and Bob, measure time and comm.
 2. Engineering Reference
 - Pick a simple setting, e.g. DH-based exact 2-party PSI
 - Test vectors, could feed into (1) above
 3. Guidance Framework
 - Help explain and manage risks and tradeoffs
 - Map real-life requirements to scheme properties

Performance Grid

Bob's Set Size →

Alice's Set Size ↓



Pick a Setting:
2-party PSI,
semi-honest,
64-bit data type

Cite the scheme
with the best time
or communication

**Covers quite a few
flavors:**

- Private Set Membership
- Asymmetric PSI
- Small Set PSI
- Symmetric PSI

Engineering Reference

- DH-based PSI has been pretty stable from [Meadows86, HFH99]
- New constructions, implementations, metrics, surveys, etc. still popularly use this as a basis [KLSAP17, IKNPRSSSY20, RT21, ...]
- Lots of RFCs already exist for DH-like protocols in non-PSI settings
- We're writing an RFC-like document for the basic construction
 - Reference implementations
 - Test vectors
 - Is this too soon?

Guidance Framework

- Is performance/functionality really the barrier to adoption?
- If potential users really wanted to compute an intersection there are other alternatives:
 - Under NDA
 - Using trusted third parties or consortiums
 - What's stopping them now, and how can PSI help bridge that gap?
- Organizations need help mapping their problems and requirements onto a solution
- Provide a guidance document/roadmap for adopting