Systematic Data Structure Lower Bounds via the Query-with-Sketch Model

Sumegha Garg, Songhua He, Yuanzhi Li, Periklis A. Papakonstantinou, Xin Yang

Rutgers U *Rutgers U* *CMU* *Rutgers U* *Meta*

Does information trade for probe complexity?

A natural question:

Given function $f: \mathcal{X}^N \to \mathcal{Y}$. Can the leakage of some nontrivial information from input $X \in \mathcal{X}^N$ reduce the probe complexity of f?

Affirmative example:

Computing $Sum = \sum_{i=1}^{n} X_i$ given $X \in \{0,1\}^n$ requires $\Omega(n)$ probes. However, if the lowest $\frac{2}{3}\log n$ bits of Sum are revealed, sampling $n^{2/3}$ computes the higher-order of Sum w.h.p.

How about its hardness?

What we do in this work

I. Formalize this model as a new query-with-sketch model and build a min-entropy tool for showing lower bounds

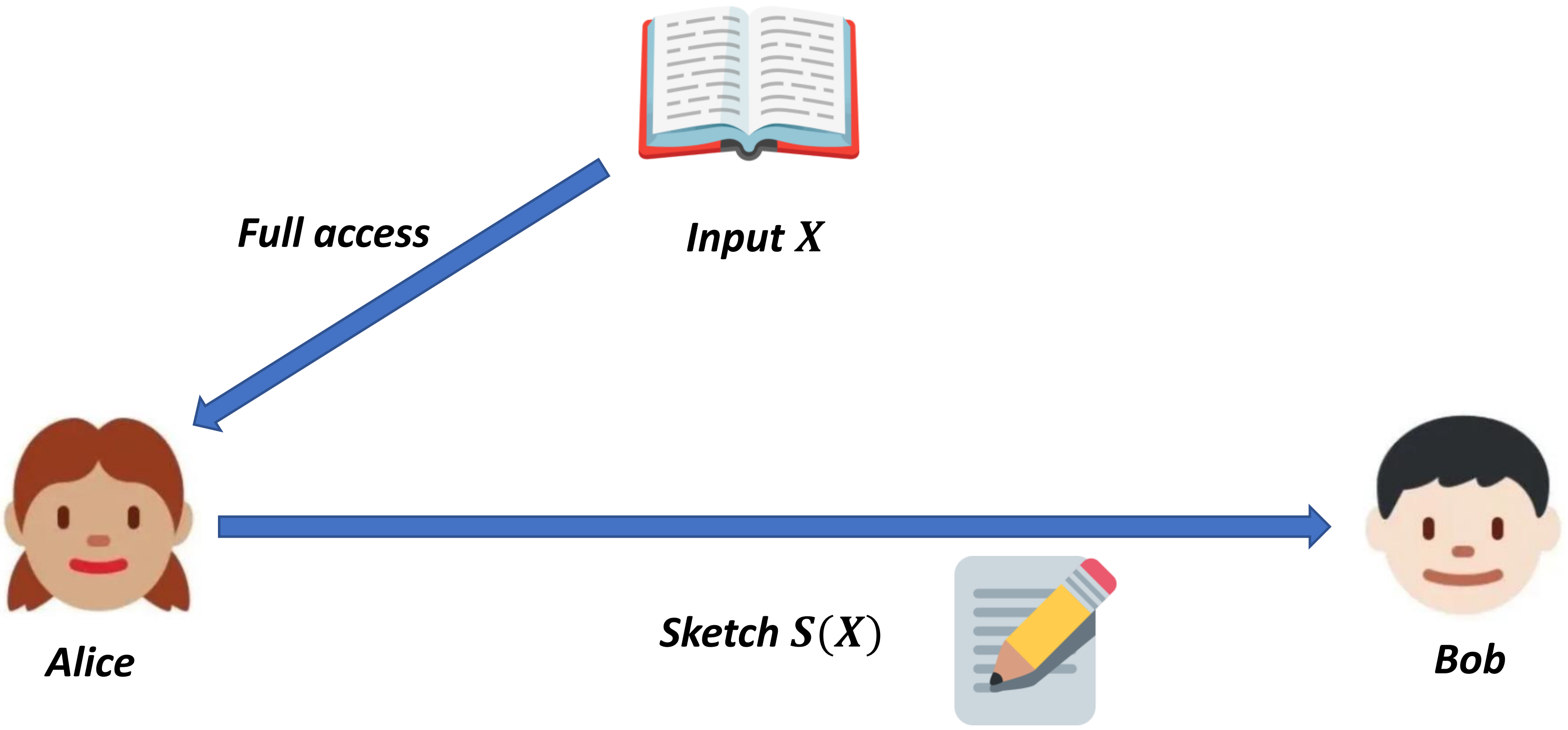
II. We generalize it to a general framework for systematic data structures

III. As applications, we give new data structure lower bounds for Approximate Matrix Powering (AMP), Set Intersection/Disjointness, (Approximate) All Pairs Shortest Paths (APSP).

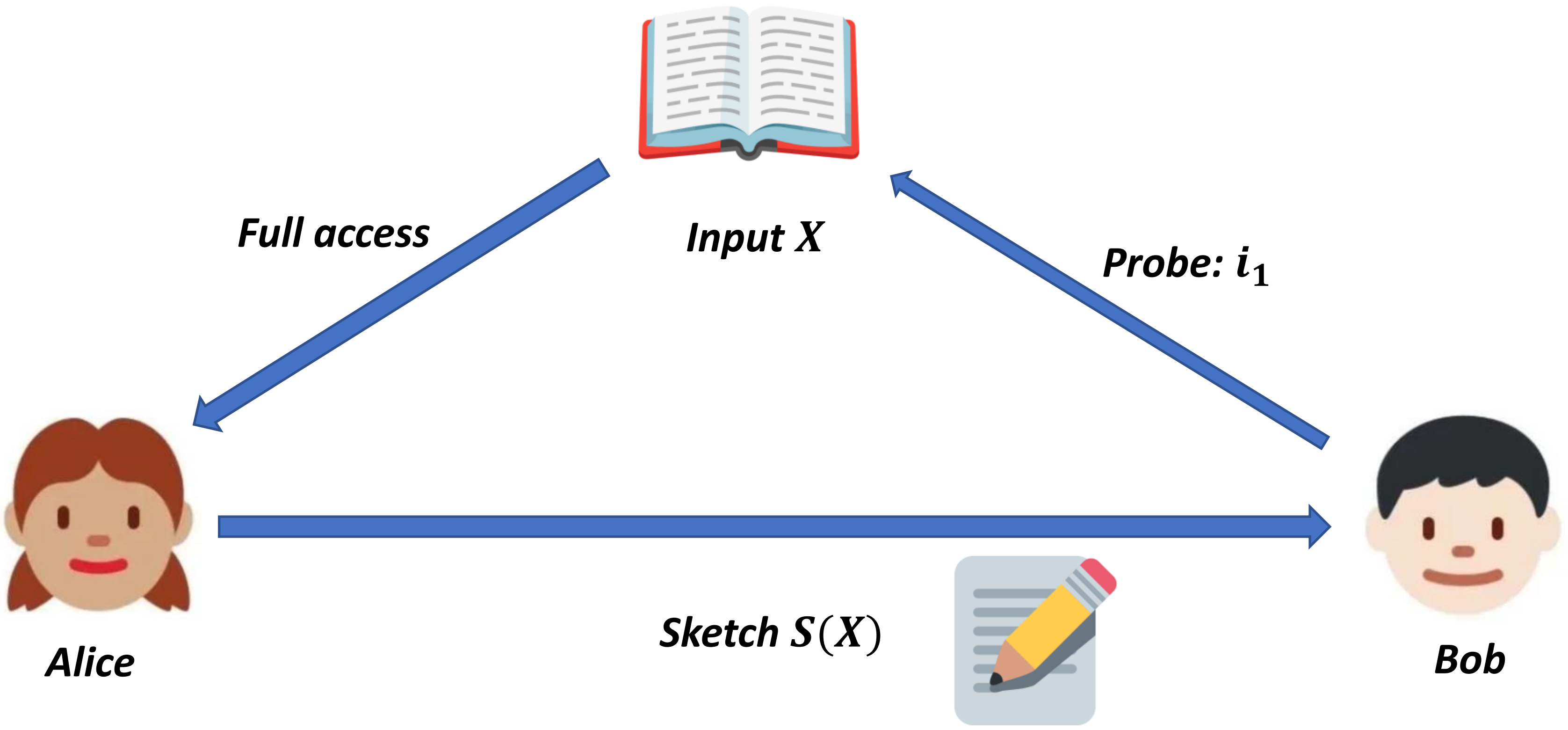
Part I

Query-with-sketch model and min-entropy lemma

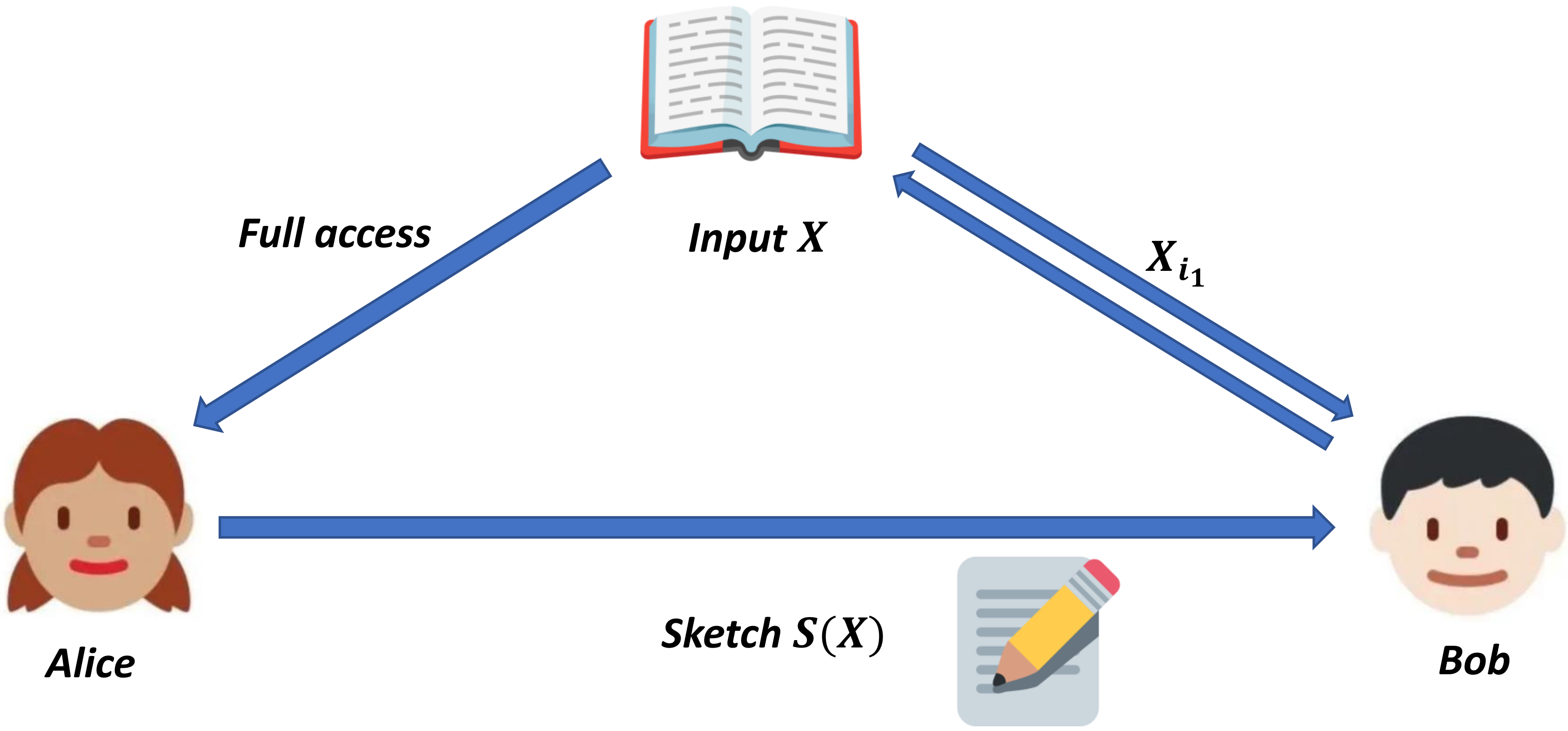
Query-with-sketch (QwS) model



Query-with-sketch (QwS) model

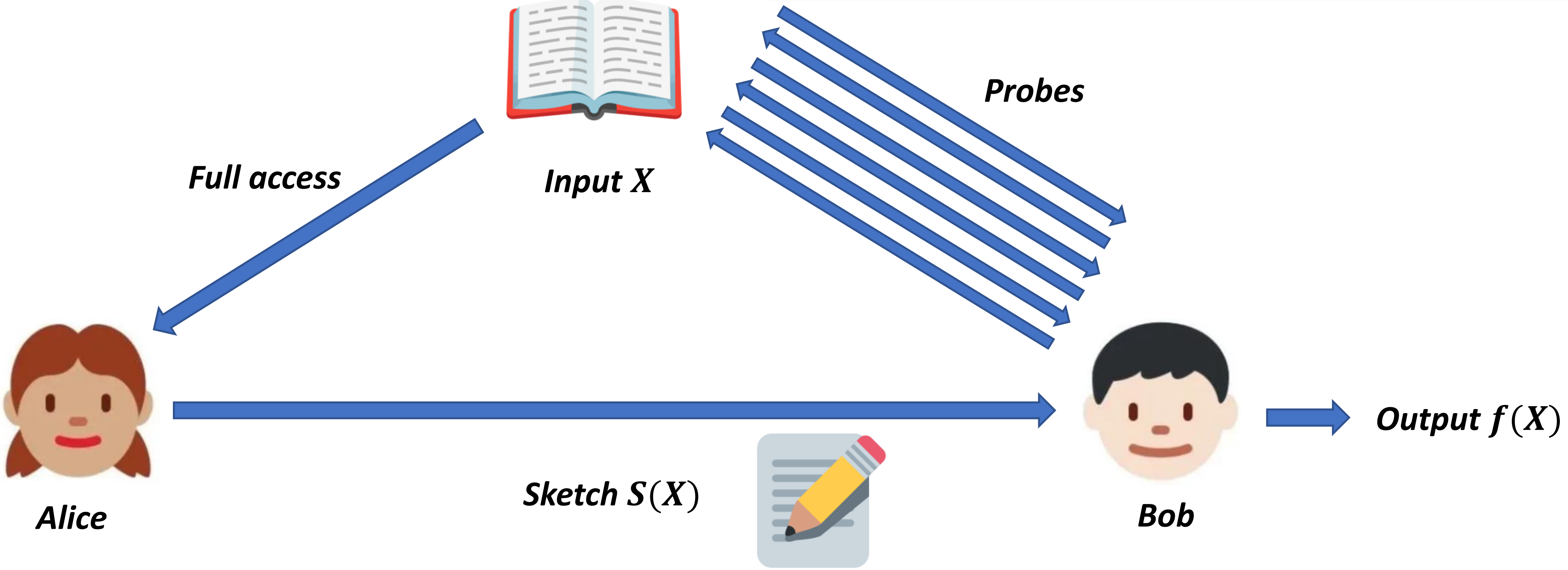


Query-with-sketch (QwS) model



Query-with-sketch (QwS) model

[NRS'98] studied this model for the direct sum of Boolean functions.



A different view of the QwS model

An algorithm for $f: \mathcal{X}^N \to \mathcal{Y}$ with probe complexity q is a decision tree of depth q, that computes f correctly on $\mathcal{X}^N$.

An algorithm for $f: \mathcal{X}^N \to \mathcal{Y}$ in the *QwS model* with probe complexity q and *sketch length* r is a *collection of* 2^r decision trees of depth q, each computing f correctly on a *subset* of $\mathcal{X}^N$.

Min-entropy lemma

High-level: if there is no sketch from Alice, and Bob always has $2r$ bits of uncertainty about the output $f(X)$ even after making q probes, then r bits of sketch cannot help.

Min-entropy lemma

Definition (min-entropy): Let $X \leftarrow \mu$. Its min-entropy is

$$H_{\min}(X) \coloneqq \min_x - \log Pr_{X \leftarrow \mu}[X = x]$$

Definition (partial assignment): A partial assignment to X of length q is an event of the form $(X_{i_1} = x_{i_1}, \ldots, X_{i_q} = x_{i_q})$.

Min-entropy lemma. Let $f: \mathcal{X}^N \rightarrow \mathcal{Y}$ and μ a distribution over $\mathcal{X}^N$. Let Π^S a deterministic algorithm for f with a sketch $S \in \{0,1\}^r$. If there is an integer q s.t. for every length-q partial assignment $\sigma = (X_{i_1} = x_{i_1}, \ldots, X_{i_q} = x_{i_q})$, $H_{\min}(f(X)|\sigma) > 2r$, then Π^S must make $> \Omega(\mathrm{q})$ queries.

Proof overview to min-entropy lemma

Condition ①: $\forall(\sigma, |\sigma| = q), H_{\min}(f(X)|\sigma) > 2r$

① $\Longrightarrow$ ② Let $X \leftarrow \mu$, *any* possible decision tree of depth q computes $f(X)$ correctly w.p. $< 2^{-2r}$ $\Longrightarrow$ ③ any QwS algorithm with sketch length r and query complexity q computes $f(X)$ correctly w.p. $< 2^{-r}$.

Proof of ① $\Longrightarrow$ ②: Fix an arbitrary decision tree T of depth q.

$$
\begin{aligned}
&\Pr_{X\leftarrow\mu}[T \text{ computes } X \text{ correctly}] \\
&\leq \textstyle\sum_{\text{leaf } \sigma} \Pr[T \text{ outputs } f(X) \mid X \text{ reaches } \sigma] \cdot \Pr[X \text{ reaches } \sigma] \\
&\leq \textstyle\sum_{\text{leaf } \sigma} 2^{-2r} \cdot \Pr[X \text{ reaches } \sigma] \\
&\leq 2^{-2r}
\end{aligned}
$$

(Because each X is assigned to a unique leaf σ)

Part II

Generalization to systematic data structures

Systematic data structures

A static data structure problem can be formalized as a function

$$f: \underbrace{\mathcal{X}^N}_{\textit{Input}} \times \underbrace{Q}_{\textit{Query}} \to \underbrace{\mathcal{Y}}_{\textit{Answer}}$$

① **Preprocessing phase**: store input X verbatim + additional $r = o(N)$ bits of information of X (redundancy).

Systematic data structures

A static data structure problem can be formalized as a function

$$f: \mathcal{X}^N \times \mathcal{Q} \to \mathcal{Y}$$

Input ***Query*** ***Answer***

② **Query phase**: a query algorithm, receiving query $\phi \in \mathcal{Q}$, with free access to the redundancy and unit-cost probe access to X, must output $f(X, \phi)$.

Why we study DS lower bounds?

- DS lower bounds have deep connections to
  - matrix-rigidity [DGW19, RR20]
  - cryptography [CGHPV20]
  - fine-grained complexity [HKNS15, CGK19]
- Systematic DS lower bounds imply circuit lower bounds [Vio19, DKKS21]

Main results

Problem 1: Approximate Matrix Powering (AMP)
Fix parameters $\alpha > 0$ and an integer $2 \leq k \leq n$
Input: A substochastic and symmetric matrix $M \in \mathbb{R}^{n \times n}$
Query: a pair $(u, v) \in [n] \times [n]$
Output: An estimation $\hat{y} \in R$ such that

$$\left|\hat{y} - M^k[u, v]\right| \leq \frac{1}{n^\alpha}$$

Theorem 1: Fix α, k. Any (randomized) systematic data structure for AMP of r bits of redundancy requires $\Omega(n^2/\mathrm{r})$ probes per query for almost all meaningful range of k and r.

Main results

Problem 2: Set Intersection
Input: n sets $A_1, A_2, \dots, A_n \subseteq [n]$
Query: Given index $(u, v) \in [n] \times [n]$
Output: the size of $A_u \cap A_v$

Theorem 2: Any randomized systematic data structure for set intersection of r bits of redundancy requires $\Omega(n^2/\mathrm{r})$ probes per query, for $r \in [\Omega(n\ log\ n), O(n^2)]$. This holds even when we only ask for $|A_u \cap A_v| \bmod 2$.

Main results

Problem 3: All Pairs Shortest Paths (APSP)
Input: an undiected graph $G = (V, E)$
Query: A pair of vertices $(u, v) \in V \times V$
Output: the shortest path length between u and v.

Theorem 3: Any randomized systematic data structure APSP of r bits of redundancy requires $\Omega(n^2/\mathrm{r})$ probes per query on average, for $r \in [\Omega(n \log n), O(n\sqrt{n}/\log^3 n)]$. This holds even when we only ask for a $(2 - \varepsilon)$-approximation of the shortest path, for every $\varepsilon > 0$.

Main results

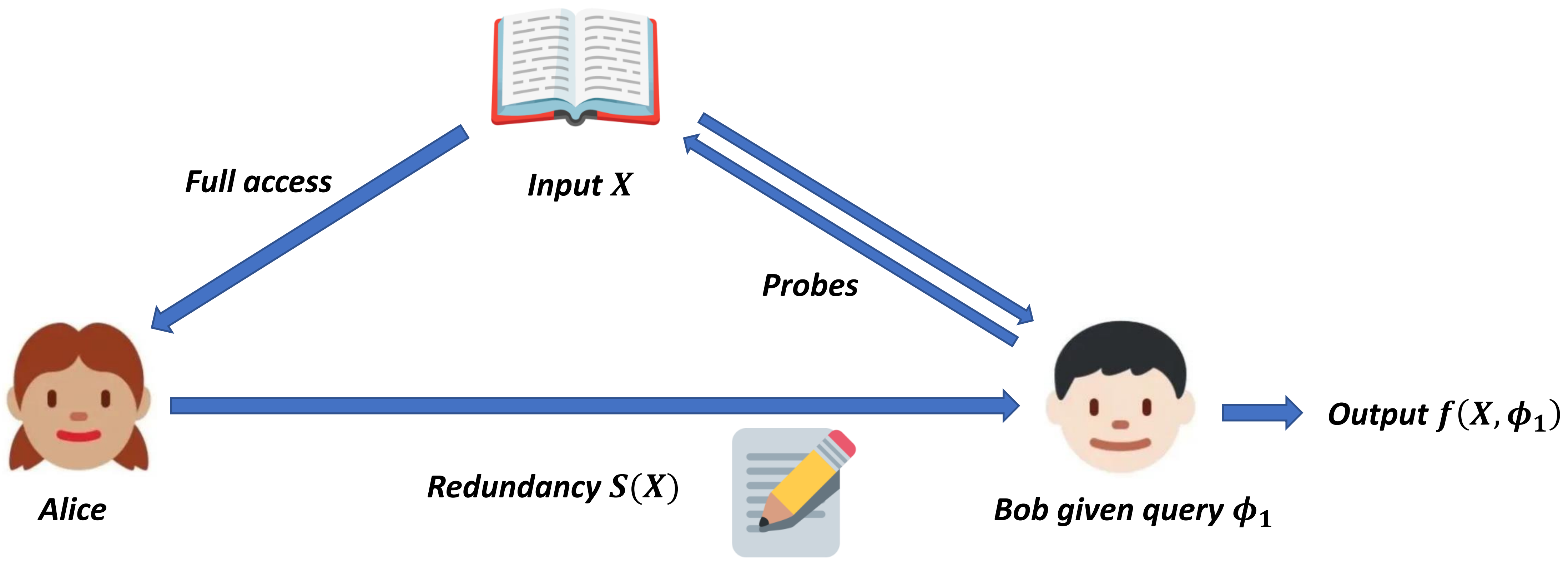
Problem 4: Set Disjointness
Input: n sets $A_1, A_2, \dots, A_n \subseteq [n]$
Query: Given index $(u, v) \in [n] \times [n]$
Output: whether $A_u \cap A_v = \emptyset$

Corollary 4: Any randomized systematic data structure for set disjointness of r bits of redundancy requires $\Omega(n^2/r)$ probes per query on average, for $r \in [\Omega(n \log n), O(n\sqrt{n}/\log^3 n)]$.

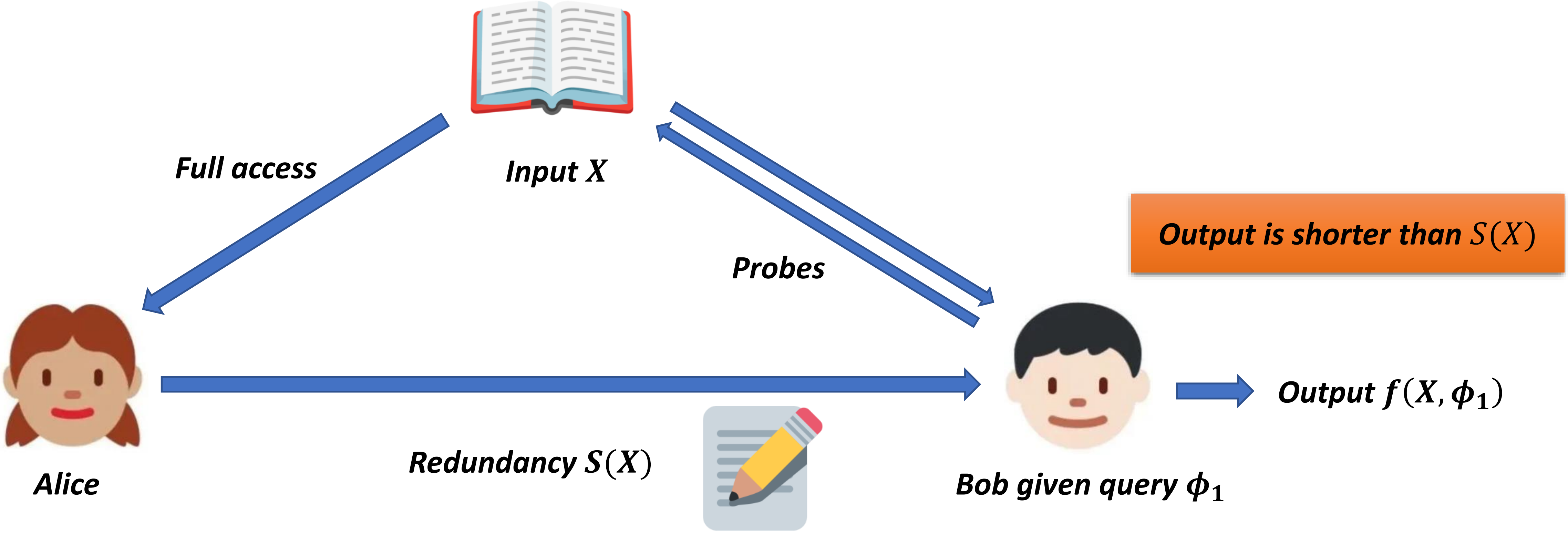
Evidence towards a conjecture [PR10]

Conjecture [PR10]: any cell-probe data structure that preprocesses n sets $A_1, \ldots, A_n \subseteq [\mathrm{polylog}(n)]$ and answers set disjointness queries using $O(1)$ probe complexity must use $\Omega(n^2)$ space.

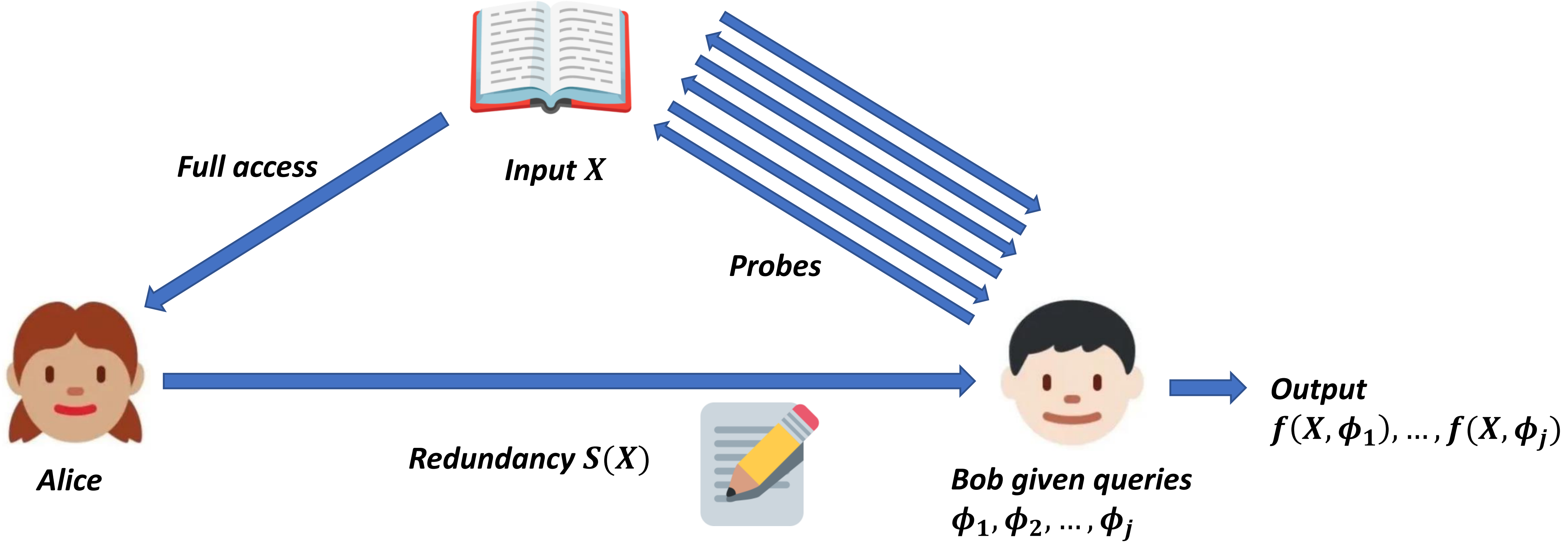
Reduction from QwS to systematic data structures



Reduction from QwS to systematic data structures



Reduction from QwS to systematic data structures



Reduction from QwS to systematic data structures

Combination of min-entropy lemma and the reduction: Given a function $f: \mathcal{X}^N \times \mathcal{Q} \to \mathcal{Y}$, for some integers q and r if there is a hard distribution μ over $\mathcal{X}^N$ and a sequence of queries $\phi_1, \dots, \phi_{10r}$ such that for any partial assignment $\sigma = (X_{i_1} = x_{i_1}, \dots, X_{i_q} = x_{i_q})$,

$$H_{\min}(f(X, \phi_1), \dots, f(X, \phi_{10r}) | \sigma) > 2r$$

Then, any systematic data structure for computing f requires $\Omega(r)$ bits of redundancy or $\Omega(q/r)$ probe complexity.

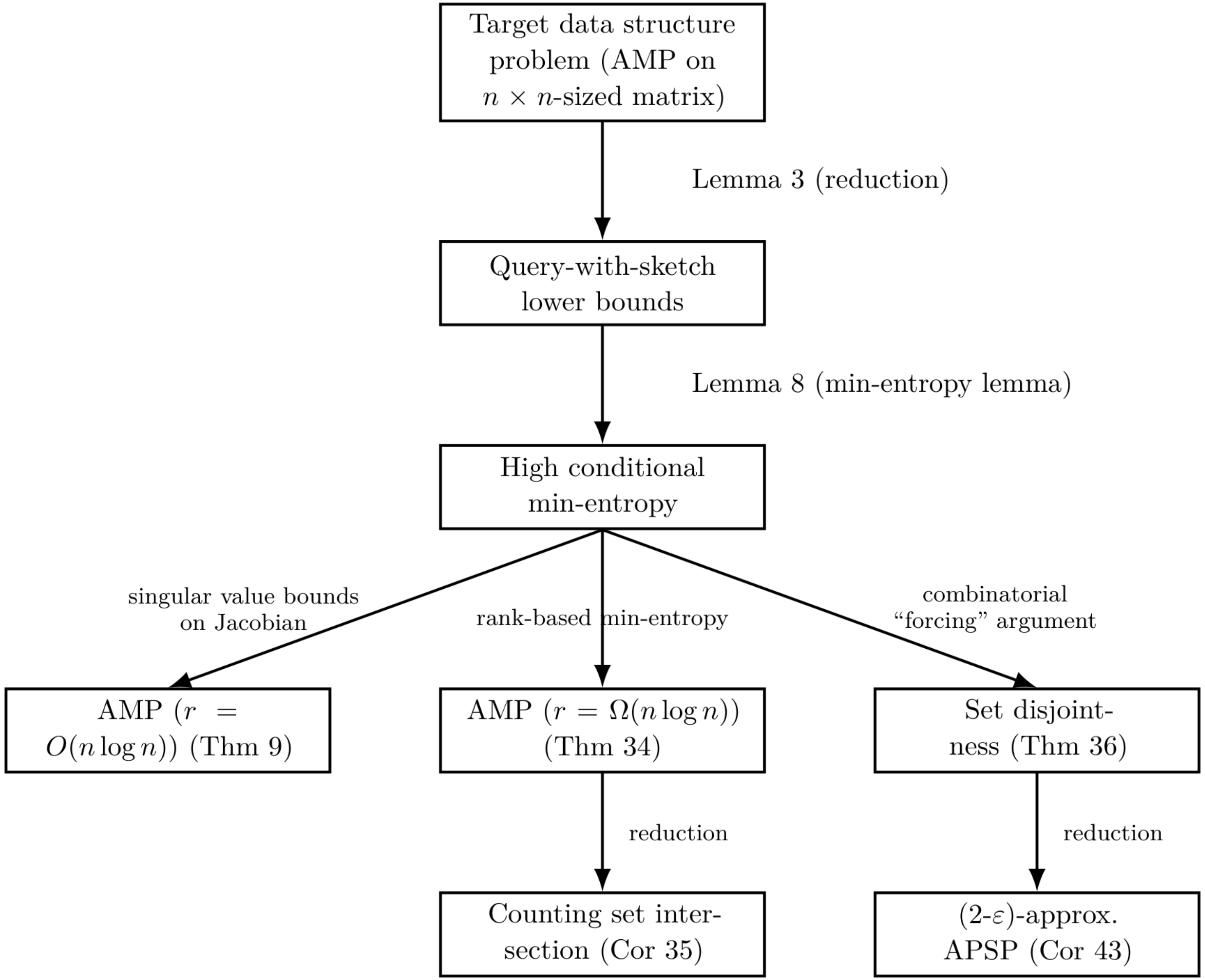


Figure 1 Our proof strategy. We reduce data structure bounds to a query-with-sketch bound (Lemma 3), which we then reduce to an information-theoretic min-entropy question (Lemma 8). We solve the latter by applying a different analytical technique for each problem.

Part III

Min-entropy bounds for Set Intersection (parity), as an example

Hard distribution μ:

Let $A_1, \ldots, A_n$ be uniformly random subsets of $[n]$.

Sequence of queries (A_u, A_v):

$$\phi_1 = \left(1, \frac{n}{2} + 1\right)$$

$$\phi_2 = \left(2, \frac{n}{2} + 2\right)$$

…

Here $r = \Omega(n \log n)$

v

		$\frac{n}{2}+1$	$\frac{n}{2}+2$	$\frac{n}{2}+3$	$\frac{n}{2}+4$	$\frac{n}{2}+5$	$\cdots$	n
	1	ϕ_1	$\phi_{\frac{n}{2}+1}$	ϕ_{n+1}				
	2		ϕ_2	$\phi_{\frac{n}{2}+2}$	ϕ_{n+2}			
	3			ϕ_3	$\phi_{\frac{n}{2}+3}$	ϕ_{n+3}		
u	4				ϕ_4	$\phi_{\frac{n}{2}+4}$	$\cdots$	
	5					ϕ_5	$\cdots$	ϕ_{10r}
	$\cdots$						$\cdots$	$\cdots$
	$\frac{n}{2}$	ϕ_n						$\phi_{\frac{n}{2}}$

To show: for every partial assignment σ of length $0.001n^2$ to the input, the conditional min-entropy of the (parity) of the output remains $> 2r$.

Proof sketch to the special case without conditioning on σ

- Write $\mathrm{M} = (\mathrm{A}_1, \dots, \mathrm{A}_{n/2}) \in \mathbb{F}_2^{n \times (n/2)}$
- M has full rank with high probability
- Fixing M, queries (u, v) with different v are independent.
- Fix one v, which appears at $10r/n$ queries $(u_1, v), \dots, (u_{10r/n}, v)$.
- The answers to these queries are exactly $\mathrm{M}_{u_1,\dots,u_{10r/n}}^T A_v$, which is uniformly random.

Open problem

- This framework cannot give lower bounds stronger than

$$\underbrace{r}_{\text{redundancy length}} \times \underbrace{(q/r)}_{\text{probe complexity}} = \Omega(\underbrace{N}_{\text{Input length}} \times \underbrace{\log|\mathcal{Y}|}_{\text{Output length}})$$

  How do we achieve stronger lower bounds?

- [CKL'18] achieved $r \cdot (q/r) = \Omega(N^{3/2})$ for OMv problem. But the proof relies on an exponentially large query set Q.