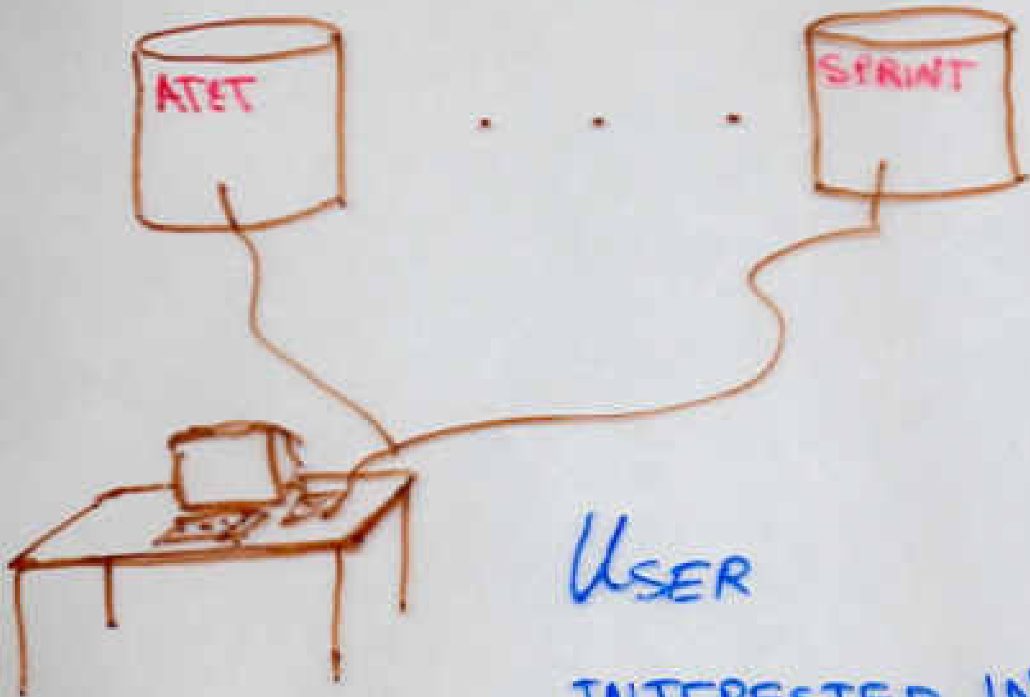
PRIVATE INFORMATION RETRIEVAL

Benny CHOR
Oded GOLDREICH
Eyal KUSHILEVITZ
Madhu SUDAN

THE PROBLEM/MODEL

- k (REPLICATED) DATABASES ($k=2.$)
  EACH HOLDING THE SAME n ITEMS



USER

INTERESTED IN ITEM $i \in [n]$

- <u>PRIVACY</u>: NO SINGLE DATABASE SHOULD KNOW i
- <u>COST</u>: COMMUNICATION (BETW' USER & DATABASES).

SOLUTIONS — PIR SCHEMES

- <u>FOR $k=2$ DATABASES</u>: COST $= 12 \cdot \sqrt[3]{n}$
- <u>$\forall\, k>2$ DATABASES</u>: COST $= O(\sqrt[k]{n})$
- FOR $k = \frac{1}{3} \cdot \log_2 n$, COST $= \frac{1}{3} \log_2^2 n \cdot \log_2 \log_2 n$

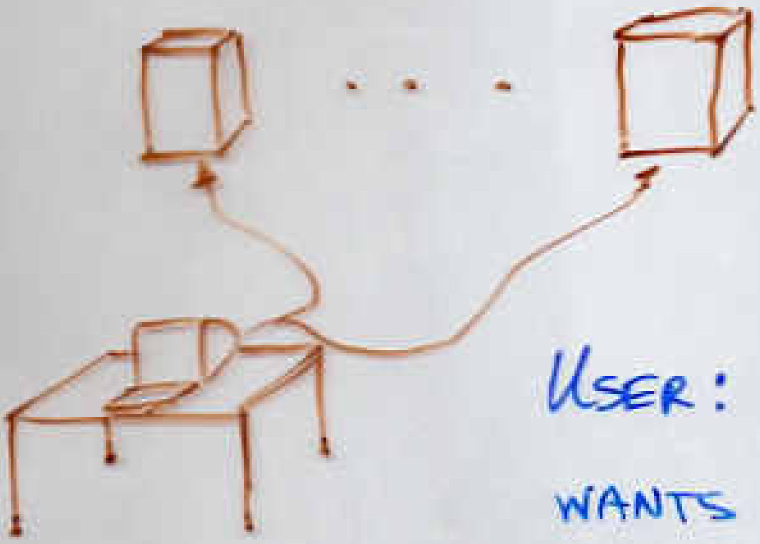
THE TRIVIAL SOLUTION ALLOW $k=1$

BUT REQUIRES COST $= n$.

AFK ≝ ABADI, FEIGENBAUM & KILIAN
BF ≝ BEAVER & FEIGENBAUM
BFKR ≝ BEAVER, FEIGENBAUM, KILIAN & ROGAW

A RELATED MODEL – "INSTANCE HIDING"

- k "POWERFUL" COMPUTERS



USER: INSTANCE $i \in \{0,1\}^{\ell}$
WANTS $f(i)$, WHERE f
IS HARD TO COMPUTE (FOR U).

- PRIVACY: AS BEFORE
- COST: USER OPERATES IN POLY($|i|$)-TIME.

RELATION TO PIR

$$[n] \longleftrightarrow \{0,1\}^{\ell} \equiv [2^{\ell}]$$

$$\text{DATABASE} \longleftrightarrow f(1)\cdot f(2) \cdots f(2^{\ell})$$

(3)

(2)

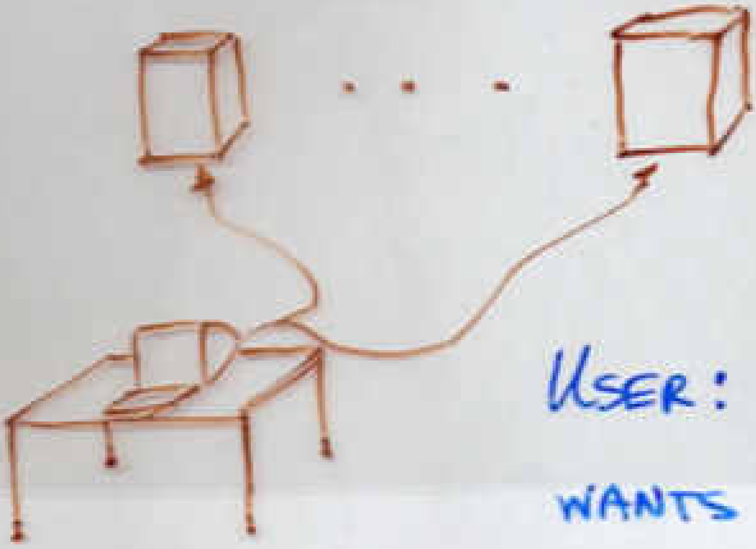
→ summary: PIR = IH scaled down
+ change of focus/complexity

– IN PIR = Po
2 FOCUS ON C
– IN I.H. = Po

AFK ≜ ABADI, FEIGENBAUM & KILIAN
BF ≜ BEAVER & FEIGENBAUM
BFKR ≜ BEAVER, FEIGENBAUM, KILIAN & ROGA

A RELATED MODEL – "INSTANCE HIDING"

- k "POWERFUL" COMPUTERS



- [AFK], $k=1$, NEG.
- [BF] $k = \ell = \log_2 n$
- [BFKR] $\forall k > 2$

USER: INSTANCE $i \in \{0,1\}^\ell$
WANTS $f(i)$, WHERE f
IS HARD TO COMPUTE (FOR U).

- PRIVACY: AS BEFORE
- COST: USER OPERATES IN POLY($|i|$)-TIME.

RELATION TO PIR

$$[n] \longleftrightarrow \{0,1\}^\ell \equiv [2^\ell]$$

$$\text{DATABASE} \longleftrightarrow f(1) \cdot f(2) \cdots f(2^\ell)$$

→ summary: PIR = IH scaled down
+ change of focus/complexity [illegible]

– IN PIR = P[illegible]
& FOCUS ON [illegible]
– IN I.H. = P[illegible]

A SIMPLE PIR FOR $k=2$

DATABASE = $x \in \{0,1\}^n$

DESIRED ITEM $i \in [n]$

USER SELECTS UNIFORMLY $S \subseteq [n]$

- SENDS S TO DATABASE_1
- SENDS $S \oplus \{i\}$ TO DATABASE_2

DATABASE, UPON RECEIVING $R \subseteq [n]$,

RETURNS $\bigoplus_{j \in R} x_j$

USER XORs THE 2 BITS.

- PRIVACY: OK. 
- COST: LINEAR IN n 

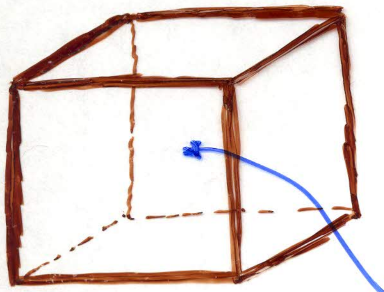
NO BETTER THAN "TRIVIAL".

So why did I waste your time?

... because this simple idea can be generalized to something useful

GENERALIZATION TO $K = 2^d$ (e.g. $d=3$)

VIEW $x = x_1, \ldots, x_n$ AS RESIDING IN A CUBE:



$$m = \sqrt[3]{n}$$

$$[n] \equiv [m] \times [m] \times [m]$$

WANTED x_{i_1, i_2, i_3}

$$i \equiv \underline{(i_1, i_2, i_3)}$$

USER SELECTS UNIFORMLY $S_1, S_2, S_3 \subseteq [m]$

- SENDS $S_1 \oplus \{\underline{i_1}\}^{\sigma_1}$, $S_2 \oplus \{\underline{i_2}\}^{\sigma_2}$, $S_3 \oplus \{\underline{i_3}\}^{\sigma_3}$

  TO DATABASE "NUMBER" $\sigma_1 \sigma_2 \sigma_3 \in \{0,1\}^3$.

DATABASE, UPON RECEIVING $R_1, R_2, R_3 \subseteq [m]$,

RETURNS $\bigoplus_{j_1 \in R_1} \bigoplus_{j_2 \in R_2} \bigoplus_{j_3 \in R_3} x_{j_1, j_2, j_3}$

USER XORs THE 8 BITS

- PRIVACY: OK. ☺
- COST: $O(m) = O(\sqrt[3]{n})$. ☺

But this is with 8 DATABASES
AND I've promised such performance with 2.

A CLOSER LOOK AT $\sqrt[3]{n}$ SOLUTION

COMMUNICATION BETWEEN USER & DATABASE

$$\text{User} \xrightarrow[= 3 \cdot m \text{ BITS}]{\text{3 SUBSETS}} \text{Database}$$

$$\xleftarrow[\text{1 (XOR) BIT}]{}$$

- SUPPOSE DB_{000} RECEIVES (R_1, R_2, R_3).
- IT KNOWS THAT DB_{100} HAS RECEIVED ONE OF THE FOLLOWING $(R_1 \oplus \{j\}, R_2, R_3)$, $j = 1, \ldots, m$.
- IT CAN "SIMULATE" DB_{100} BY REPLYING WITH ALL m POSSIBILITIES. (m BITS!)

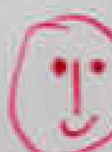
$\Rightarrow$ DB_{000} SIMULATES $DB_{100}, DB_{010}, DB_{001}$

& DB_{111} SIMULATES $DB_{011}, DB_{101}, DB_{110}$

$\Rightarrow$ 2 DATABASE SCHEME WITH

PRIVACY: OK.

COST: $O(\sqrt[3]{n})$.



$U \xrightarrow{3m} DB_{000}$

$U \xleftarrow{3m+1} DB_{000}$

IN GENERAL ($d \geq 3$)

- A PIR FOR 2^d DATABASES WITH COMMUNICATION

USER $\xrightarrow{d \text{ SUBSETS OF } [\sqrt[d]{n}]}$ DATABASE

$= d \cdot \sqrt[d]{n}$ BITS

$\xleftarrow{\hspace{3cm}}$

1 (XOR) BIT

- A COVERING CODE (OF RADIUS 1) FOR $\{0,1\}^d$

USING K CODEWORDS. $\left(\frac{2^d}{d+1} \leq k \leq 2^d \right)$

"VOLUME BOUND"

$\Rightarrow$ A PIR FOR K DATABASES

WITH COMMUNICATION COST $\propto 2^d \cdot d \cdot \sqrt[d]{n}$.

d	2^d	K	VOLUME BOUND.	
3	8	2	2	PERFECT
4	16	4	4	
5	32	7	6	
6	64	12	10	
7	128	16	16	PERFECT

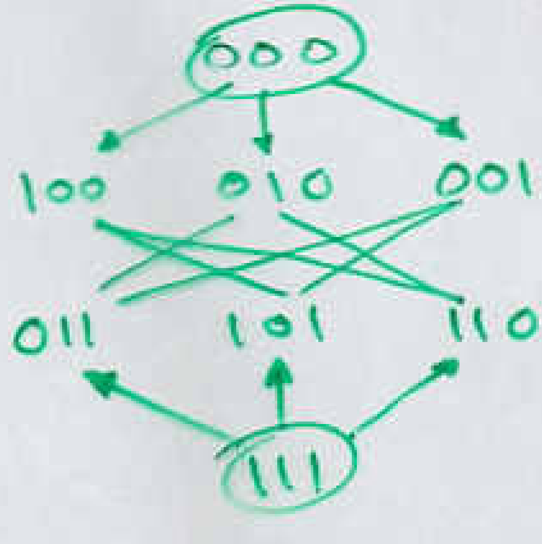
COVERING CODES (OF RAD' 1)

$C \subseteq \{0,1\}^d$ IS A <u>COVERING CODE</u> FOR $\{0,1\}^d$ IF

$\forall\, x \in \{0,1\}^d \;\exists\, y \in C$ S.T. $\mathrm{DIST}(x,y) \le 1$.

HAMMING DISTANCE.

E.G. $\{000, 111\}$ IS A COVERING CODE FOR $\{0,1\}^3$.



VOLUME BOUND $\longrightarrow$ EACH CODEWORD COVERS $d+1$ STRINGS

$\Longrightarrow$ $\#\text{CODEWORDS} \ge \dfrac{2^d}{d+1}$

EQUALITY FOR <u>PERFECT COVER</u>

POLYNOMIAL INTERPOLATION PIRs

$x \in \{0,1\}^n \implies \mathcal{X} : [n] \to \{0,1\}$

APPLICATION: [BF]

$r_1, \ldots, r_q \in_R F$ $\quad (q \in \{\ell, m\})$

$f(t) \triangleq \hat{\mathcal{X}}(i_1 + t \cdot r_1, \ldots, i_q + t \cdot r_q)$

WHERE $(i_1, \ldots, i_q)$ IS THE ITEM SOUGHT.

OBTAIN FROM j^{TH} DATABASE THE VALUE $f(j)$

INTERPOLATE TO OBTAIN $f(0) = \mathcal{X}(i_1, \ldots, i_q)$.

POLYNOMIAL INTERPOLATION PIRs

$x \in \{0,1\}^n \implies X:[n]\to\{0,1\}$

$$\mathrel{|||}$$

$X:\binom{m}{d}\to\{0,1\}$, $\quad m \approx \sqrt[d]{n}$

$\approx$[BFKR]

$\hat{X}: F^m \to F$

DEGREE d MULTI-LIN' EXT'

I.E., $\hat{X}(z_1,\dots,z_m) \triangleq \sum_{S\in\binom{m}{d}} \left(\prod_{j\in S} z_j\right)\cdot X(S)$.

APPLICATION: [BF]

$r_1,\dots,r_q \in_R F \qquad (q\in\{\ell, m\})$

$f(t) \triangleq \hat{X}(i_1+t\cdot r_1,\dots,i_q+t\cdot r_q)$

WHERE $(i_1,\dots,i_q)$ IS THE ITEM SOUGHT.

OBTAIN FROM j^{TH} DATABASE THE VALUE $f(j)$

INTERPOLATE TO OBTAIN $f(0)=\hat{X}(i_1,\dots,i_q)$.

OPEN PROBLEM

BEST PIR SCHEMES

#DB	COST		CONJ.
2	$O(\sqrt[3]{n})$		$\Omega(\sqrt[3]{n})$
$k > 2$	$O(\sqrt[k]{n})$	$\xleftrightarrow[\text{TIGHT}]{\text{NOT}}$	$\Omega(\sqrt[\text{[illegible]}]{n})$

"BEST" LOWER BOUND

FOR $K=2$, IF USER IS ONLY ALLOWED A SINGLE BOOLEAN QUERY THEN $|\text{QUERY}| = \Omega(n)$.

ADDENDUM (JULY 2008)

The conjectured lowerbound for multiple-server PIRs(i.e., the case of $k > 2$) was disproved a couple of years afterwards by Ambainis in his paper ``An Upper Bound On The Communication Complexity of Private Information Retrieval'' (24th ICALP, LNCS 1256, pages 401--407, 1997).Ambainis established an upper bound of $n^{1/(2k-1)}$, which became my revised conjecture for a tight result.

This conjecture was disproved by Beimel, Ishai, Kushilevitz, and Raymond in their paper ``Breaking the $O(n^{1/(2k-1)})$ barrier for information-theoretic private information retrieval''(43rd FOCS, pages 261--270, 2002).

My last attempt at a conjecture was that for every k there exists a constant $c = c(k)$ such that a k-server PIR requires communication complexity n^c. Furthermore, I expected $c(k)$ to equal the reciprocal of some small polynomial.

This last conjecture seems to be disproven by Yekhanin in his paper``Towards 3-Query Locally Decodable Codes of Subexponential Length'' (39th STOC, pages 266--274, 2007).

I dare not make further conjectures....