

Lecture 13

Private Information Retrieval

Stefan Dziembowski
University of Rome
La Sapienza



SAPIENZA
UNIVERSITÀ DI ROMA

BiSS 2009
Bertinoro International
Spring School
2-6 March 2009



Plan

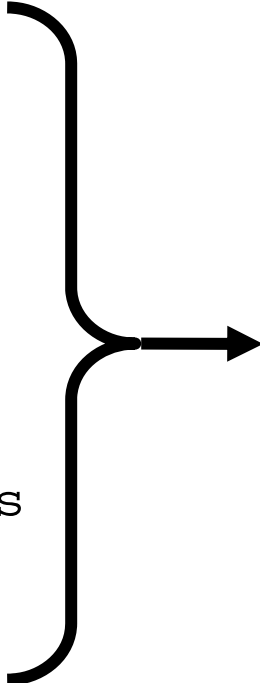


1. Motivation and definition
2. Information-theoretic impossibility
3. A construction of Kushilevitz and Ostrovsky
4. Overview of some other related results

AOL search data scandal (2006)

#4417749:

- clothes for age 60
- 60 single men
- best retirement city
- jarrett arnold
- jack t. arnold
- jaylene and jarrett arnold
- gwinnett county yellow pages
- rescue of older dogs
- movies for dogs
- sinus infection



Thelma Arnold
62-year-old widow
Lilburn, Georgia

Observation

The owners of databases know a lot about the users!

This poses a risk to users' privacy.

E.g. consider database with stock prices...

Can we do something about it?



problematic

We can:

- **trust** them that they will protect our secrecy,
- or
- use **cryptology**! ←

How can crypto help?



Note: this problem has nothing to do with secure communication!

Our settings



user **U**



database **D**

A new primitive:

Private Information Retrieval (PIR)

Plan

1. Definition of PIR
2. An ideal PIR doesn't exist
3. Construction of a computational PIR
4. Open problems

Literature:

- B. Chor, E. Kushilevitz, O. Goldreich and M. Sudan,
Private Information Retrieval, Journal of ACM, 1998
- E. Kushilevitz and R. Ostrovsky
**Replication Is NOT Needed: SINGLE Database,
Computationally-Private Information Retrieval**, FOCS 1997

Question

How to protect privacy of queries?



user **U**

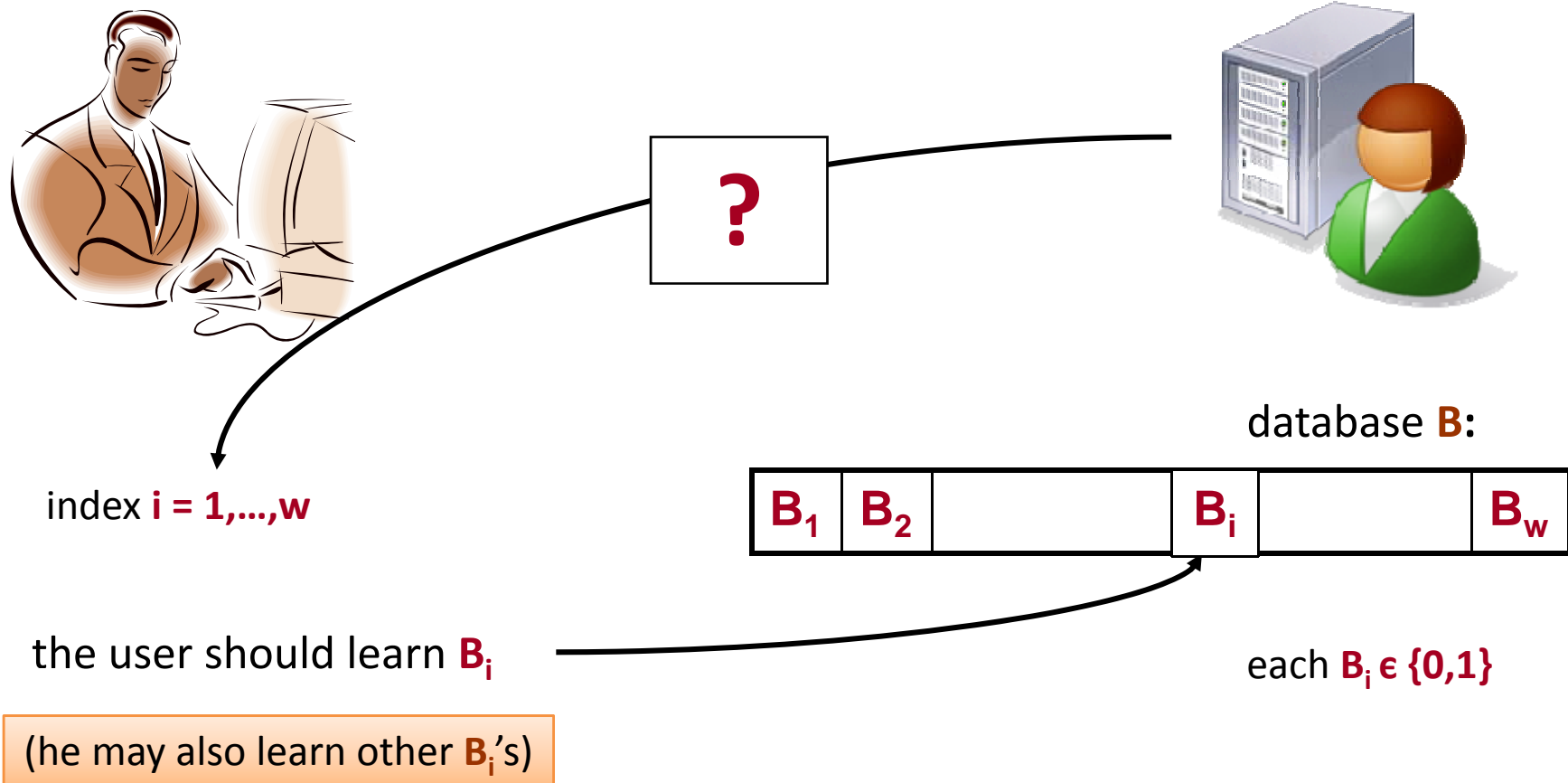
wants to retrieve some
data from **D**



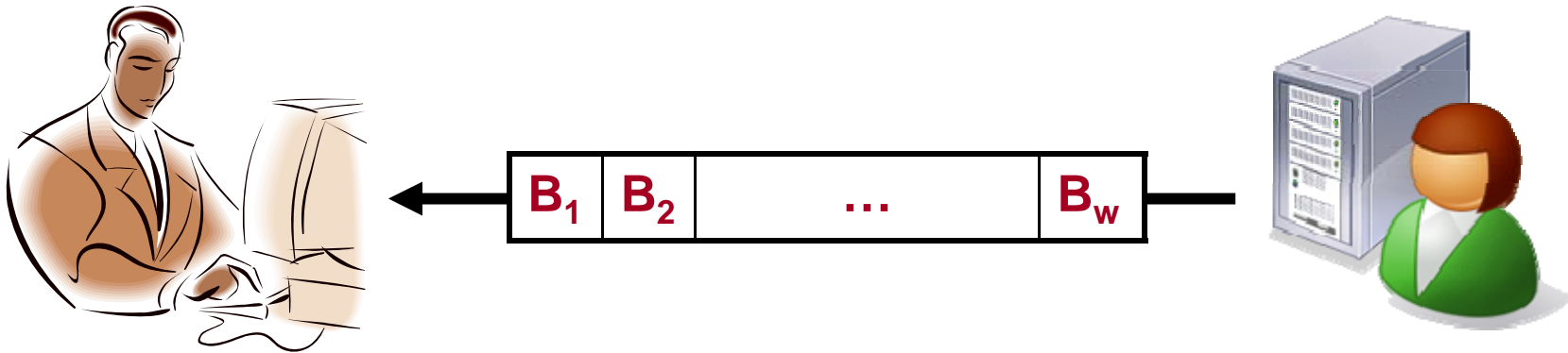
database **D**

shouldn't learn what **U**
retrieved

Let's make things simple!



Trivial solution



The database simply sends everything to the user!

Non-triviality

The previous solution has a drawback:

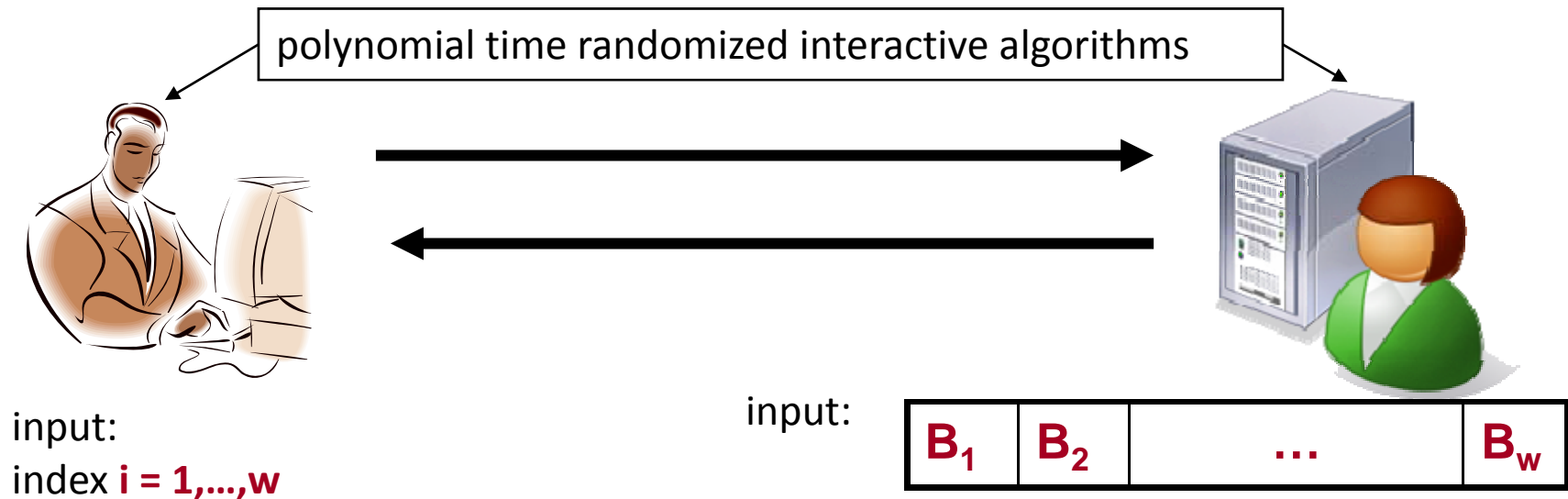
the communication complexity is huge!

Therefore we introduce the following requirement:

“Non-triviality”:

**the number of bits communicated between U and D
has to be smaller than w .**

Private Information Retrieval (PIR)



This property needs to be defined more formally

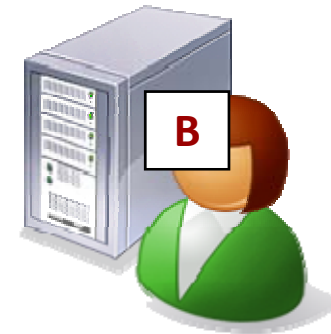
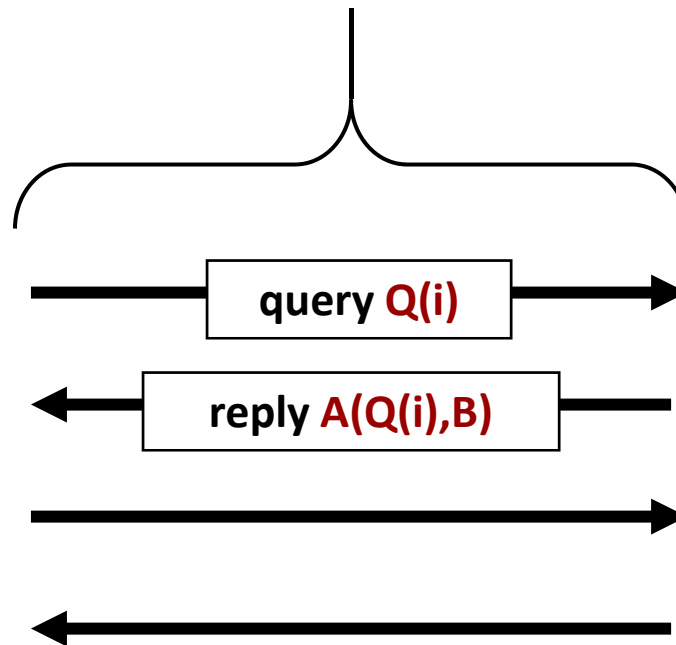
- at the end the user learns B_i ← correctness
- the database does not learn i ← secrecy (of the user)
- the total communication is $< w$ ← non-triviality

Note: secrecy of the database is not required

How to define secrecy of the user [1/2]?

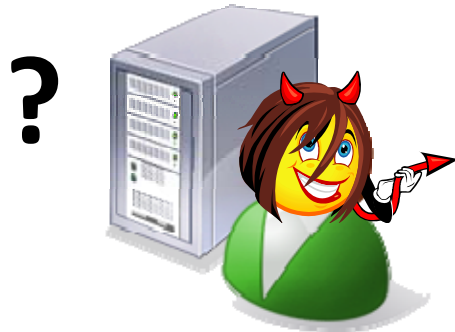
Def. $T(i,B)$ – transcript of the conversation.

For fixed i and B
 $T(i,B)$
is a **random variable**
(since the parties are
randomized)



How to define secrecy of the user [2/2]?

Secrecy of the user: for every $i, j \in \{0,1\}$



single-round case:

it is impossible to distinguish between $Q(i)$ and $Q(j)$

multi-round case:

it is impossible to distinguish between $T(i,B)$ and $T(j,B)$

even if the adversary is malicious

For simplicity say that for any i and j the distributions of $T(i,B)$ and $T(j,B)$ have to be identical

Plan

1. Motivation and definition
2. Information-theoretic impossibility
3. A construction of Kushilevitz and Ostrovsky
4. Overview of some other related results



PIR doesn't exist [2/4]

Observation:

secrecy \rightarrow

if

T' is possible for some B and i

then

it is possible for B and all the other i 's

databases B

T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'
T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'

indices i

PIR doesn't exist [3/4]

non-triviality \rightarrow $\text{length}(\text{transcript}) < \text{length}(\text{database})$



$\# \text{ transcripts} < \# \text{ databases}$



there has to exist T' that is possible for
two databases B_0 and B_1

	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'		$\leftarrow B_0$
	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'		$\leftarrow B_1$

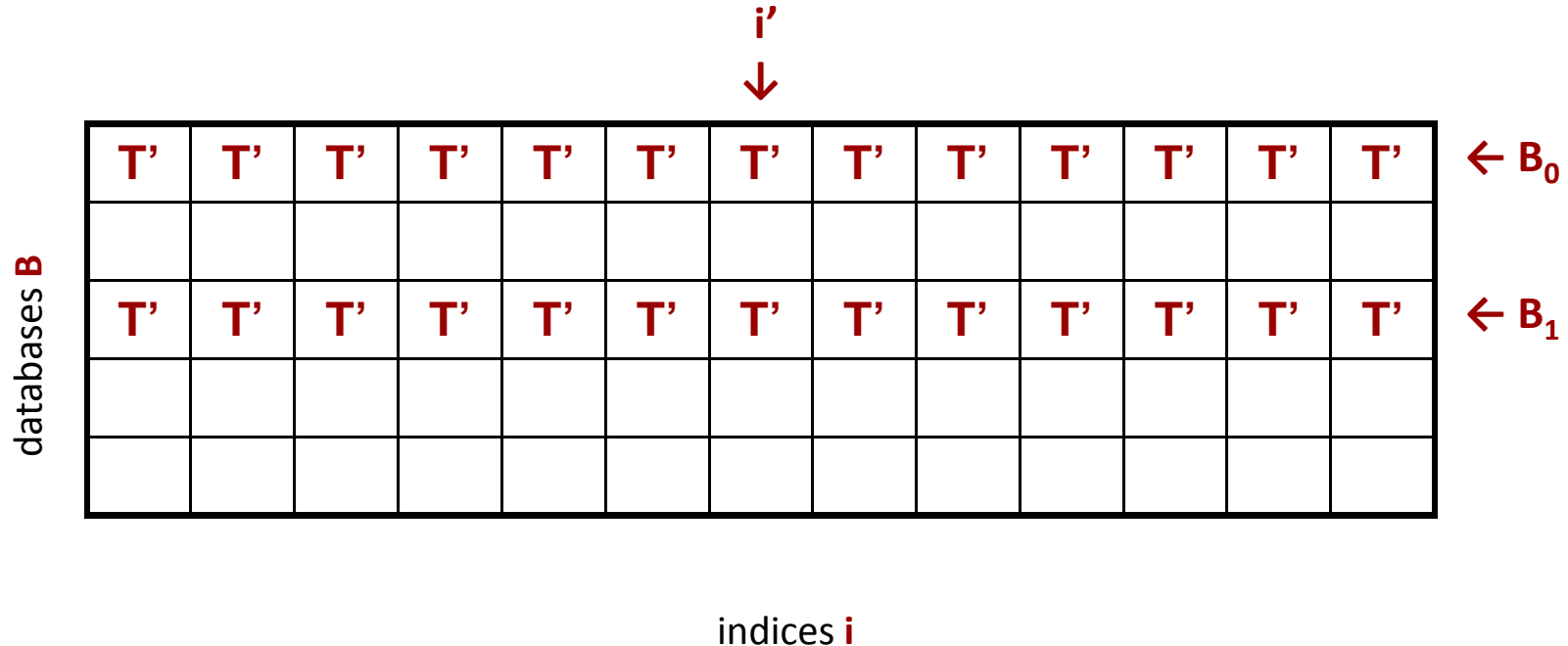
databases B

indices i

PIR doesn't exist [4/4]

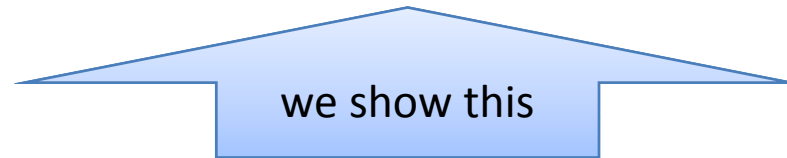
B_0 and B_1 differ on at least one index i'
So, if i' is the input of the user then

correctness \rightarrow contradiction



So PIR doesn't exist!

- How to bypass the impossibility result?
- **Two ideas:**
 - limit the computing power of a cheating database



- use a larger number of “independent” databases

Plan

1. Motivation and definition
2. Information-theoretic impossibility
3. A construction of Kushilevitz and Ostrovsky
4. Overview of some other related results



Computationally-secure PIR

computational-secrecy:

?



For every $i, j \in \{0,1\}$

it is impossible to distinguish

efficiently

between

$T(i,B)$ and $T(j,B)$

Formally: for every **polynomial-time** probabilistic algorithm **A** the value:

$$| P(A(T(i,B)) = 0) - P(A(T(j,B))=0) |$$

should be **negligible**.

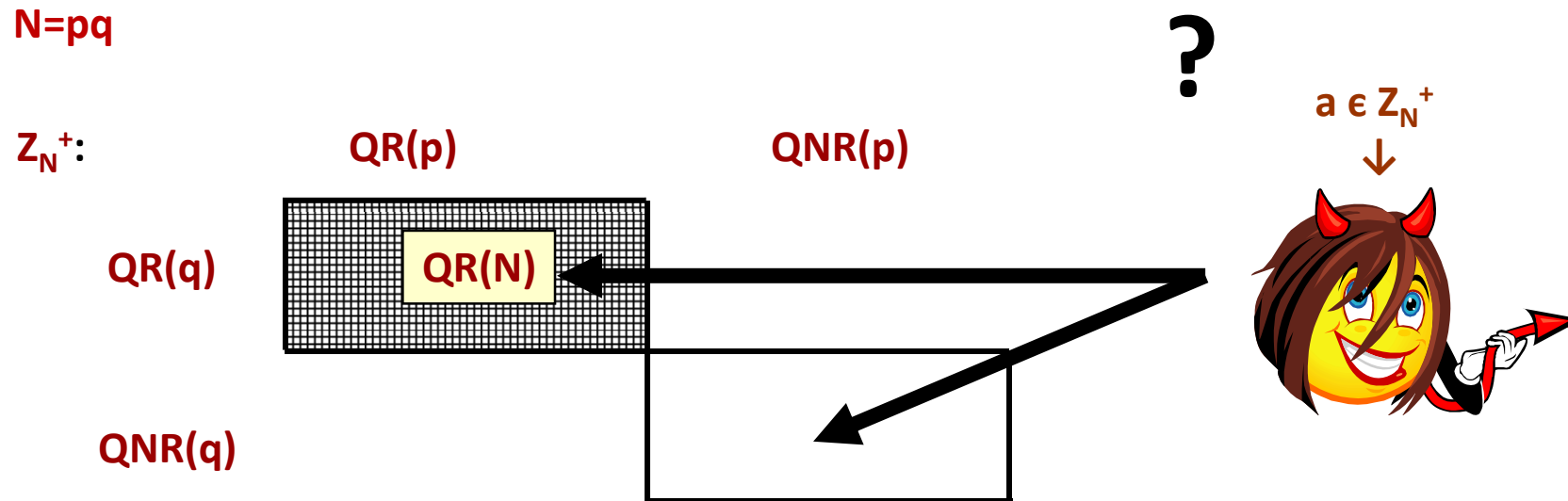
Hardness assumptions?

Kushilevitz and R. Ostrovsky **Replication Is NOT Needed: SINGLE Database, Computationally-Private Information Retrieval**, FOCS 1997

construct PIR based on the

Quadratic Residuosity Assumption

Quadratic Residuosity Assumption (QRA)



Quadratic Residuosity Assumption (QRA):

For a random $a \in Z_N^+$ it is computationally hard to determine if $a \in QR(N)$.

Formally: for every **polynomial-time** probabilistic algorithm G the value:

$$|P(G(a) = Q(a)) - 0.5|$$

(where a is random) is **negligible**.

Homomorphism of $QR(pq)$

$$Q(N,a) := \begin{cases} 1 & \text{if } a \in QR(N) \\ 0 & \text{otherwise} \end{cases}$$

Homomorphism: for all $a, b \in \mathbb{Z}_N^+$

$$Q(N,ab) = Q(N,a) \text{ xor } Q(N,b)$$

We are ready to construct PIR!

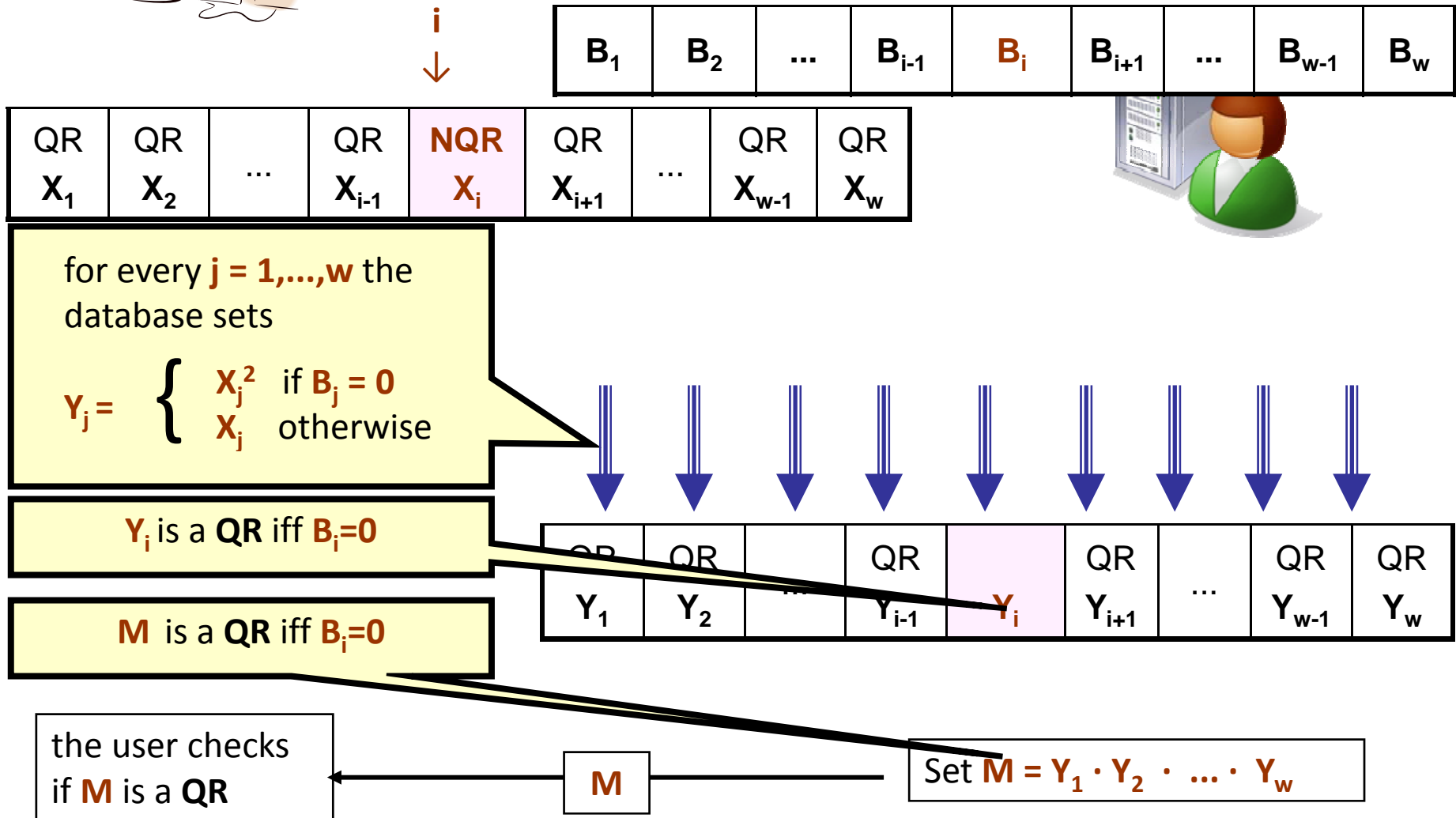
Our PIR will work in the group \mathbb{Z}_N^+ , where $N=pq$.

What's so good about this group?:

- testing membership in \mathbb{Z}_N^+ is easy,
- testing membership in $QR(N)$ is hard for random elements on \mathbb{Z}_N^+ , unless one knows p and q .
- homomorphism of Q !



First (wrong) idea



Problems!

PIR from the previous slide:

- **correctness** ✓
- **security?**

To learn **i** the database would need to distinguish **NQR** from **QR**. ✓

QR	QR	...	QR	NQR	QR	...	QR	QR
x_1	x_2	...	x_{i-1}	x_i	x_{i+1}	...	x_{w-1}	x_w



- **non-triviality?** doesn't hold!

communication:

user \rightarrow database: $|B| \cdot |Z_n^*|$

database \rightarrow user: $|Z_n^*|$

Call it:
 $(|B|, 1)$ - PIR

How to fix it?

Idea:

Given:

$(|\mathbf{B}|, 1)$ -PIR.

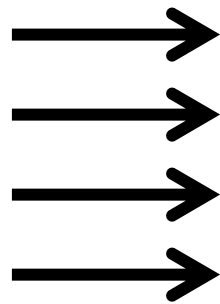
construct

$(\sqrt{|\mathbf{B}|}, \sqrt{|\mathbf{B}|})$ -PIR.

Suppose that $|\mathbf{B}| = v^2$ and present \mathbf{B} as a $\mathbf{v} \times \mathbf{v}$ -matrix:

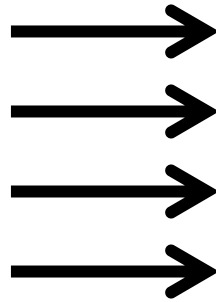
B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16
----	----	----	----	----	----	----	----	----	-----	-----	-----	-----	-----	-----	-----

consider each
row as a
separate
database



An improved idea

execute \mathbf{v}
 $(\mathbf{v}, 1)$ - PIRs
in parallel



B1	B2	B3	B4
B5	B6	B7	B8
B9	B10	B11	B12
B13	B14	B15	B16

A 4x4 grid of elements labeled B1 through B16. A bracket above the grid is labeled \mathbf{v} . A bracket to the right of the grid is labeled \mathbf{v} .

Looks even worse:
communication:
user \rightarrow database: $\mathbf{v}^2 \cdot |\mathbb{Z}_n^*|$
database \rightarrow user: $\mathbf{v} \cdot |\mathbb{Z}_n^*|$

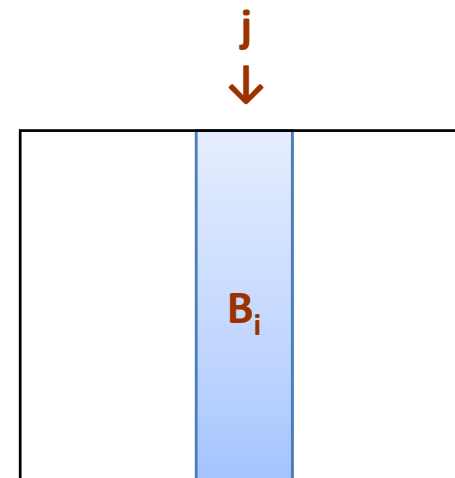
The method

Let j be the column where B_i is.

In every “row” the user asks for the j th element

So, instead of sending \mathbf{v} queries the user can send one!

Observe: in this way the user learns all the elements in the j th column!



So we are done!

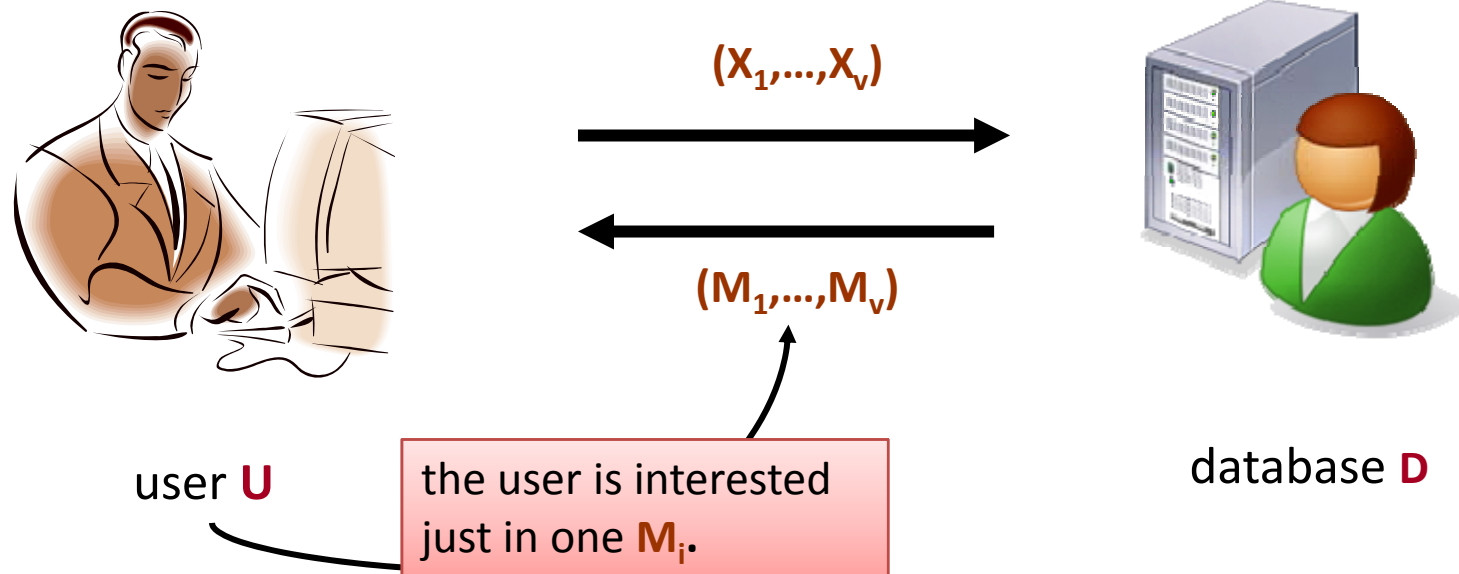
PIR from the previous slide:

- **correctness** ✓
- **non-triviality:**
communication complexity = $2v|B| \cdot |Z_n| v$
- **security?**
The to learn i the database would need to distinguish NQR from QR.

Formally:

from
any adversary that **breaks our scheme**
we can construct
an algorithm that **breaks QRA**

Improvements



Idea: apply PIR recursively!

Plan

1. Motivation and definition
2. Information-theoretic impossibility
3. A construction of Kushilevitz and Ostrovsky
4. Overview of some other related results



Complexity of PIRs – overview of the results

Communication:

- “recursive” PIR of [KO97]:
for every c : $O(|B|^c)$
- [Cachin, Micali, Stadler, 1999]:
poly-logarithmic in $|B|$
- [Lipmaa, 2005]:
 $O(\log^2 |B|)$

For practical analysis see:

- [Sion, Carbunar]
On the Computational Practicality of Private Information Retrieval.

their conclusion:

It is the time-complexity that matters

In **real-life:**

it is still more practical
to **transmit the entire database.**



Extensions

- Symmetric PIR (also protect privacy of the database).
[**Gertner, Ishai, Kushilevitz, Malkin. 1998**]
- Searching by key-words
[**Chor, Gilboa, Naor, 1997**]
- Public-key encryption with key-word search
[**Boneh, Di Crescenzo, Ostrovsky, Persiano**]

©2009 by Stefan Dziembowski. Permission to make digital or hard copies of part or all of this material is currently granted without fee *provided that copies are made only for personal or classroom use, are not distributed for profit or commercial advantage, and that new copies bear this notice and the full citation.*