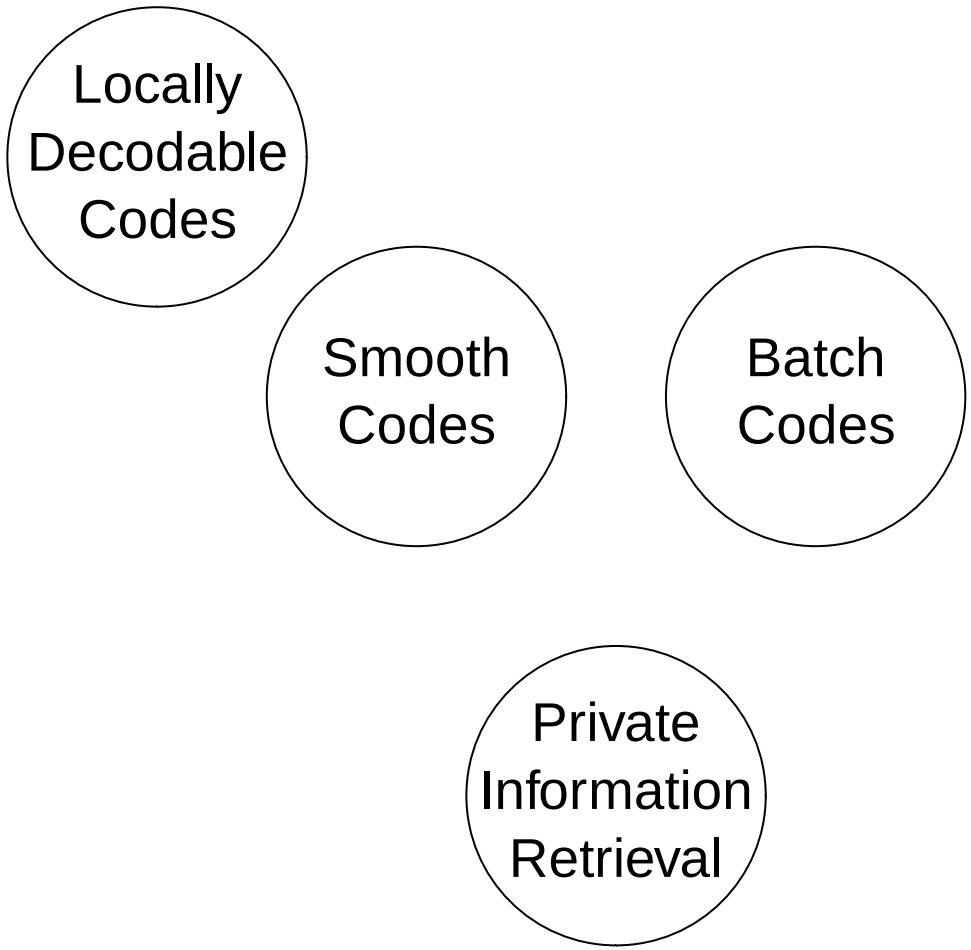
Codes and Privacy

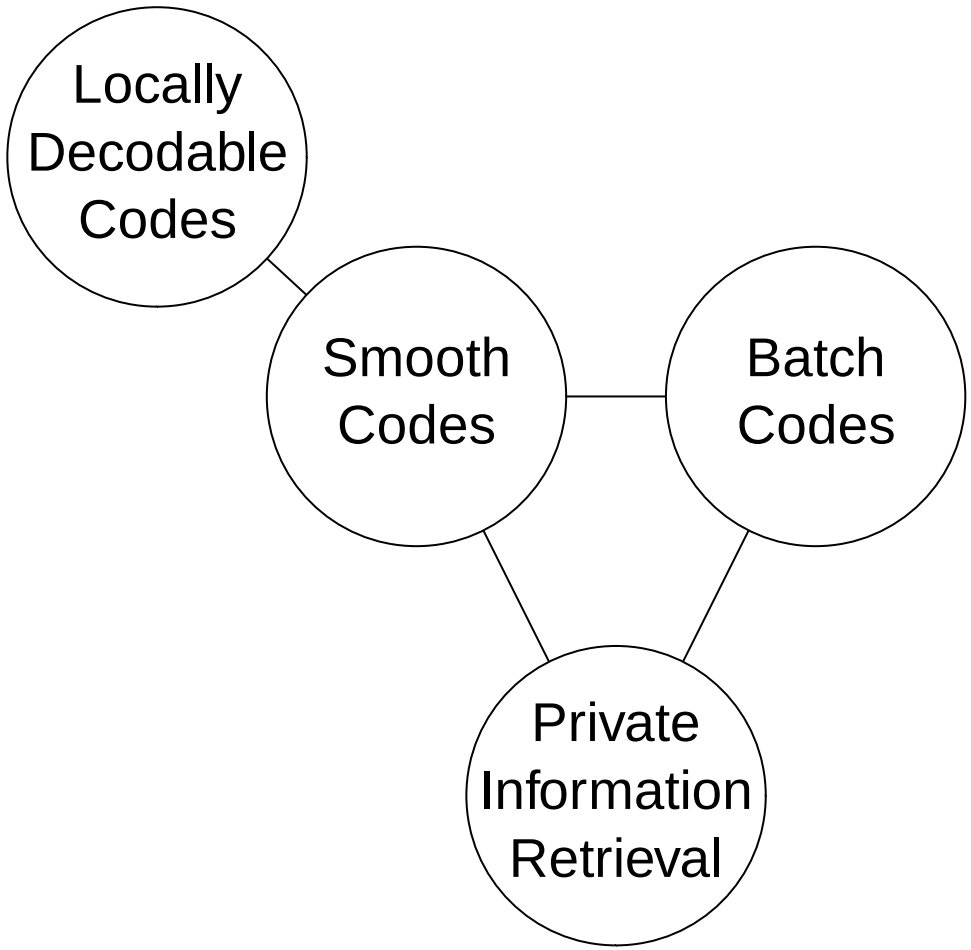
Maria Bras-Amorós

Universitat Rovira i Virgili

Interconsolider i-MATH & ARES
Castro Urdiales
15 de Mayo, 2008

Contents


Locally Decodable Codes
Smooth Codes
Batch Codes
Private Information Retrieval

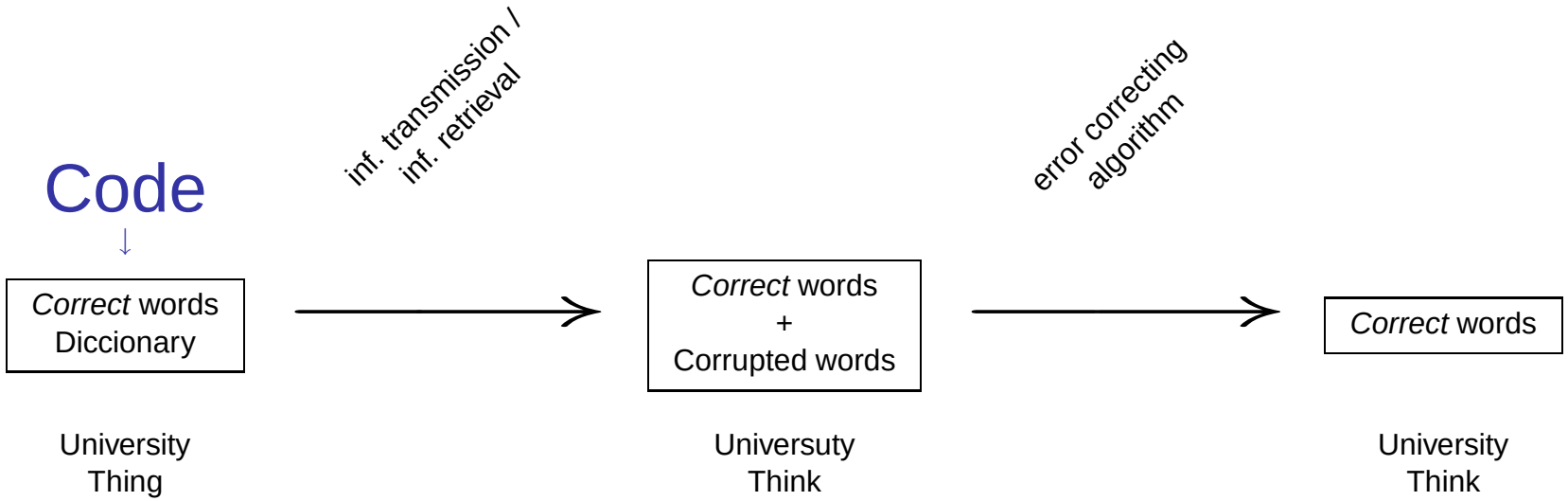

Locally Decodable Codes
Smooth Codes
Batch Codes
Private Information Retrieval

Let Σ be an alphabet.
Let $x \in \Sigma^n$ be a database.
Suppose x is replicated among one or more servers.

A PIR protocol allows a user to retrieve x_i from the server(s) while hiding i.

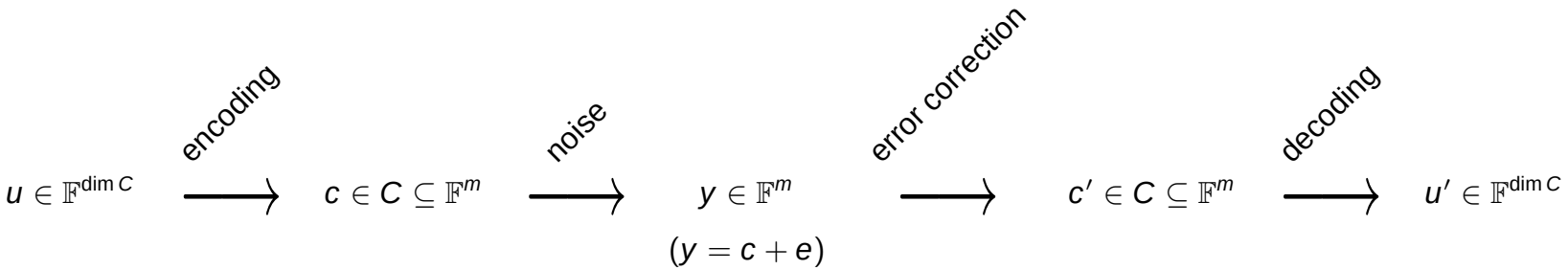
- Information-theoretic PIR:
  - more than one server,
  - privacy refers to each individual server.
- Computational PIR:
  - one server which is computationally bounded,
  - privacy is relaxed to *computational* privacy.

Codes and errors



Linear codes

A linear code of length m is a vector subspace C of $\mathbb{F}^m$.



Usual error correction approach:

obtain u' from y.

Locally decoding approach:

obtain the bit u'_i from a (small) number of randomly chosen bits in y.

Example

The Hamming code $H_2(7,4)$ has parity check matrix

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Its dual code has encoding function

$$\begin{array}{rcl} E : \mathbb{Z}_2^3 & \to & \mathbb{Z}_2^7 \\ (x_1, x_2, x_3) & \mapsto & (x_1, x_2, x_3, x_1 + x_2, x_1 + x_3, x_2 + x_3, x_1 + x_2 + x_3) \end{array}$$

$$(\; x_1, \; x_2, \; x_3 \;) \mapsto (\; x_1, \; x_2, \; x_3, \; x_1 + x_2, \; x_1 + x_3, \; x_2 + x_3, \; x_1 + x_2 + x_3 \;)$$

Probabilistic locally decoding algorithm D

Input: $y \in \mathbb{Z}_2^7$
$i \in \{1, 2, 3\}$

- Randomly take $j_1 \in \{1, \ldots, 7\}$.
- If $j_1 \neq i$
  - Take the only j_2 such that $E(x)_{j_1} + E(x)_{j_2} = x_i$.
  - Output $y_{j_1} + y_{j_2}$.
- If $j_1 = i$
  - Output y_{j_1}.

$$\begin{array}{ccccccccccccccc} (& x_1, & x_2, & x_3 &) & \mapsto & (& x_1, & x_2, & x_3, & x_1+x_2, & x_1+x_3, & x_2+x_3, & x_1+x_2+x_3 &) \\ (& 1, & 0, & 1 &) & \mapsto & (& 1, & 0, & 1, & 1, & 0, & 1, & 0 &) \end{array}$$

Probabilistic locally decoding algorithm D

Input: $y \in \mathbb{Z}_2^7$
$i \in \{1,2,3\}$

- Randomly take $j_1 \in \{1,\ldots,7\}$.
- If $j_1 \neq i$
  - Take the only j_2 such that $E(x)_{j_1} + E(x)_{j_2} = x_i$.
  - Output $y_{j_1} + y_{j_2}$.
- If $j_1 = i$
  - Output y_{j_1}.

$$\begin{array}{ccccccccccccccc} (& x_1, & x_2, & x_3 &) & \mapsto & (& x_1, & x_2, & x_3, & x_1+x_2, & x_1+x_3, & x_2+x_3, & x_1+x_2+x_3 &) \\ (& 1, & 0, & 1 &) & \mapsto & (& 1, & 0, & 1, & 1, & 0, & 1, & 0 &) \end{array}$$

Probabilistic locally decoding algorithm D

Input: $y \in \mathbb{Z}_2^7$ $\qquad (1,0,1,1,0,1,0)$
$i \in \{1,2,3\}$ $\qquad 1$

- Randomly take $j_1 \in \{1,\ldots,7\}$.
- If $j_1 \neq i$
  - Take the only j_2 such that $E(x)_{j_1} + E(x)_{j_2} = x_i$.
  - Output $y_{j_1} + y_{j_2}$.
- If $j_1 = i$
  - Output y_{j_1}.

$$\begin{array}{ccccccccccccccc} (& x_1, & x_2, & x_3 &) & \mapsto & (& x_1, & x_2, & x_3, & x_1+x_2, & x_1+x_3, & x_2+x_3, & x_1+x_2+x_3 &) \\ (& 1, & 0, & 1 &) & \mapsto & (& 1, & 0, & 1, & 1, & 0, & 1, & 0 &) \end{array}$$

Probabilistic locally decoding algorithm D

Input: $y \in \mathbb{Z}_2^7$ $(1,0,1,1,0,1,0)$
$i \in \{1,2,3\}$ 1

- Randomly take $j_1 \in \{1,\ldots,7\}$. 4
- If $j_1 \neq i$
  - Take the only j_2 such that $E(x)_{j_1} + E(x)_{j_2} = x_i$.
  - Output $y_{j_1} + y_{j_2}$.
- If $j_1 = i$
  - Output y_{j_1}.

$$\begin{array}{ccccccccccccccc} (& x_1, & x_2, & x_3 &) & \mapsto & (& x_1, & x_2, & x_3, & x_1+x_2, & x_1+x_3, & x_2+x_3, & x_1+x_2+x_3 &) \\ (& 1, & 0, & 1 &) & \mapsto & (& 1, & 0, & 1, & 1, & 0, & 1, & 0 &) \end{array}$$

Probabilistic locally decoding algorithm D

Input: $y \in \mathbb{Z}_2^7$ $(1,0,1,1,0,1,0)$
$i \in \{1,2,3\}$ 1

- Randomly take $j_1 \in \{1,\ldots,7\}$. 4
- If $j_1 \neq i$
  - Take the only j_2 such that $E(x)_{j_1} + E(x)_{j_2} = x_i$. 2
  - Output $y_{j_1} + y_{j_2}$. $y_4 + y_2 = 1$
- If $j_1 = i$
  - Output y_{j_1}.

$$\begin{array}{ccccccccccccccc} (& x_1, & x_2, & x_3 &) & \mapsto & (& x_1, & x_2, & x_3, & x_1+x_2, & x_1+x_3, & x_2+x_3, & x_1+x_2+x_3 &) \\ (& 1, & 0, & 1 &) & \mapsto & (& 1, & 0, & 1, & 1, & 0, & 1, & 0 &) \end{array}$$

Probabilistic locally decoding algorithm D

Input: $y \in \mathbb{Z}_2^7$ $(1,0,1,1,0,1,0)$
$i \in \{1,2,3\}$ 1

- Randomly take $j_1 \in \{1,\ldots,7\}$. 1
- If $j_1 \neq i$
  - Take the only j_2 such that $E(x)_{j_1} + E(x)_{j_2} = x_i$.
  - Output $y_{j_1} + y_{j_2}$.
- If $j_1 = i$
  - Output y_{j_1}.

$$\begin{array}{ccccccccccccccc} (& x_1, & x_2, & x_3 &) & \mapsto & (& x_1, & x_2, & x_3, & x_1+x_2, & x_1+x_3, & x_2+x_3, & x_1+x_2+x_3 &) \\ (& 1, & 0, & 1 &) & \mapsto & (& 1, & 0, & 1, & 1, & 0, & 1, & 0 &) \end{array}$$

Probabilistic locally decoding algorithm D

Input: $y \in \mathbb{Z}_2^7$ $(1,0,1,1,0,1,0)$
$i \in \{1,2,3\}$ 1

- Randomly take $j_1 \in \{1,\ldots,7\}$. 1
- If $j_1 \neq i$
  - Take the only j_2 such that $E(x)_{j_1} + E(x)_{j_2} = x_i$.
  - Output $y_{j_1} + y_{j_2}$.
- If $j_1 = i$
  - Output y_{j_1}. $y_1 = 1$

$$(\; x_1, \; x_2, \; x_3 \;) \mapsto (\; x_1, \; x_2, \; x_3, \; x_1 + x_2, \; x_1 + x_3, \; x_2 + x_3, \; x_1 + x_2 + x_3 \;)$$

Probabilistic locally decoding algorithm D

Input: $y \in \mathbb{Z}_2^7$
$i \in \{1, 2, 3\}$

- Randomly take $j_1 \in \{1, \ldots, 7\}$.
- If $j_1 \neq i$
  - Take the only j_2 such that $E(x)_{j_1} + E(x)_{j_2} = x_i$.
  - Output $y_{j_1} + y_{j_2}$.
- If $j_1 = i$
  - Output y_{j_1}.

Algorithm D reads at most 2 indices of y.

If $d(y, E(x)) = 0$ then for all i

$$Pr[D(y, i) = x_i] = 1.$$

If $d(y, E(x)) = 1$??
If $i = 1$ and $y - E(x) = 1000000$ then

$$Pr[D(y, i) = x_i] = \frac{6}{7}.$$

The same happens if $i = 2$ and $y - E(x) = 0100000$ or $i = 3$ and $y - E(x) = 0010000$.

In the remaining cases

$$Pr[D(y, i) = x_i] = \frac{5}{7}.$$

Definition (Katz, Trevisan, 2000)

For fixed $\delta, \epsilon \in \mathbb{R}$ and $q \in \mathbb{Z}^+$, we say that

$$E : \{0,1\}^n \to \Sigma^m$$

is a (q, δ, ϵ)-locally decodable encoding if there exists a probabilistic algorithm D such that

- For every $x \in \{0,1\}^n$, for every $y \in \Sigma^m$, with $d(y, E(x)) \leqslant \delta m$ and for all $i \in [n]$, we have

$$Pr[D(y,i) = x_i] \geqslant \frac{1}{2} + \epsilon.$$

- In every invocation, D reads at most q indices of y.

Example

In the previous example, E is a $(q = 2, \epsilon = \frac{5}{7} - \frac{1}{2} = \frac{3}{14}, \delta = \frac{1}{7})$-locally decodable encoding.

Example (trivial)

A linear code with dimension k and correction capability $t = \lfloor \frac{d-1}{2} \rfloor$ is

- $(m, t/m, 1/2)$-locally decodable code,
- $(k, 0, 1/2)$-locally decodable code.

Shortcoming

Suppose $Pr[D(\cdot, i) \text{ reads index } j] \approx 1$.
An adversary can corrupt $E(x)_j$ affecting the performance of D.

In the previous example,

$$Pr[D(\cdot, i) \text{ reads index } j] = \begin{cases} 1/7 & \text{if } i = j \\ 2/7 & \text{if } i \neq j \end{cases}$$

Definition (Katz, Trevisan, 2000)

For fixed $c, \epsilon \in \mathbb{R}$ and $q \in \mathbb{Z}^+$, we say that

$$E : \{0,1\}^n \to \Sigma^m$$

is a (q, c, ϵ)-smooth encoding if there exists a probabilistic algorithm D such that

- For every $x \in \{0,1\}^n$ and for all $i \in [n]$, we have

$$Pr[D(E(x), i) = x_i] \geqslant \frac{1}{2} + \epsilon.$$

- In every invocation, D reads at most q indices of y.
- For every $i \in [n]$ and $j \in [m]$, we have:

$$Pr[D(\cdot, i) \text{ reads index } j] \leqslant \frac{c}{m}.$$

Example

The encoding of the previous example is a $(q = 2, c = 2, \epsilon = \frac{3}{14})$-smooth encoding.

Example (trivial)

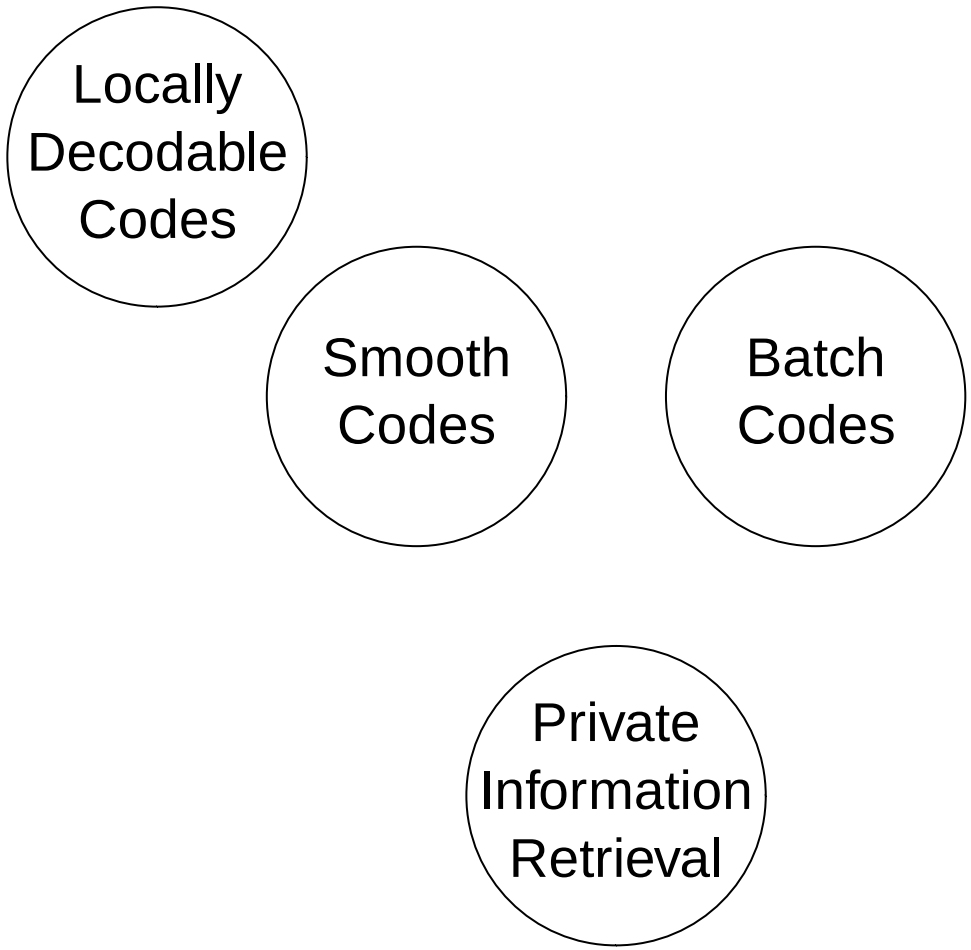
A linear code in $\mathbb{F}^m$ is a $(m, m, 1/2)$-locally decodable code.

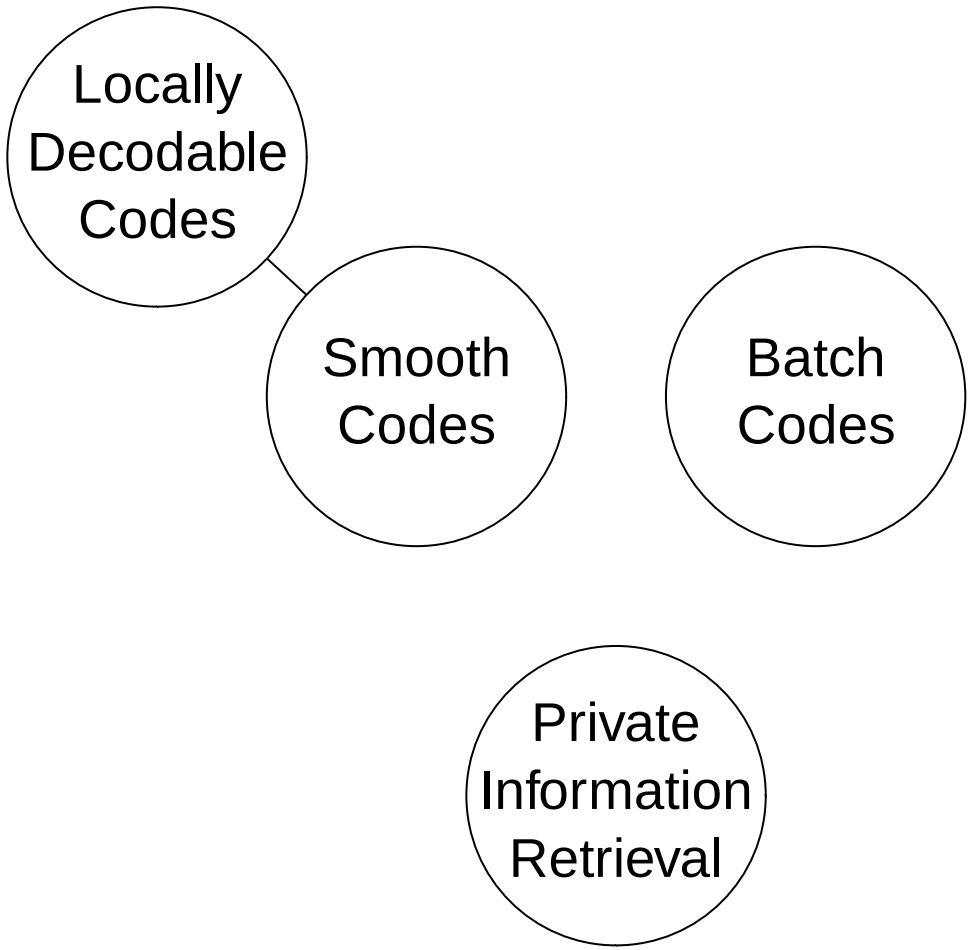
Remark

- Any (q, c, ϵ)-smooth encoding is a $(q, 0, \epsilon)$-locally decodable encoding and so, it is a (q, δ, ϵ)-locally decodable encoding for some $\delta \geqslant 0$.
- Any (q, δ, ϵ)-locally decodable encoding is a (q, m, ϵ)-smooth encoding and so it is a (q, c, ϵ)-smooth encoding for some $c \leqslant m$.

Theorem (Katz, Trevisan, 2000)

Any (q, δ, ϵ)-locally decodable encoding is a $(q, q/\delta, \epsilon)$-smooth encoding.

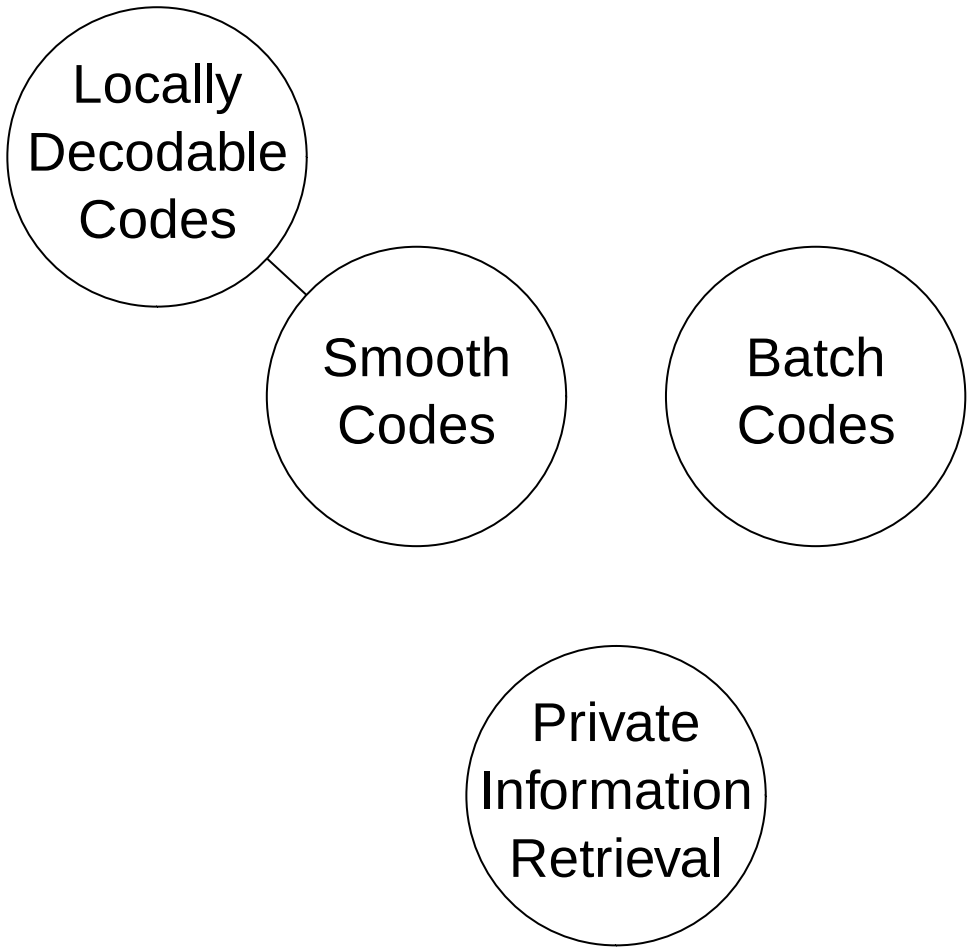

Locally Decodable Codes
Smooth Codes
Batch Codes
Private Information Retrieval

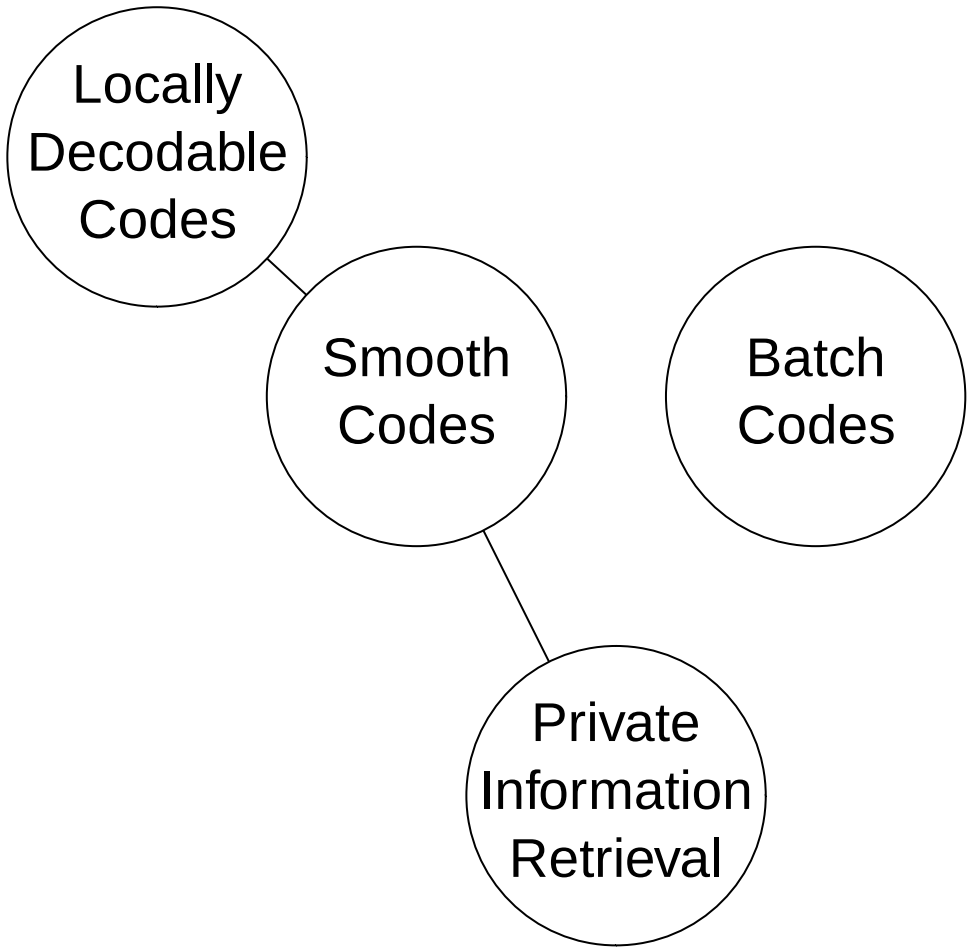

Locally Decodable Codes
Smooth Codes
Batch Codes
Private Information Retrieval

Results from Katz, Trevisan, 2000

Any (q, c, ϵ)-smooth decoding algorithm (with $c > q$) can be converted into a $(q, q, \epsilon^2/2c)$-smooth decoding algorithm with uniformly distributed queries.
A (q, q, ϵ)-smooth encoding $C : \{0,1\}^n \to \Sigma^m$ with uniformly distributed queries gives a information-theoretic PIR scheme with q servers, query size $\log m$, and answer size $\log |\Sigma|$, such that the user has probability $1/2 + \epsilon$ of correctly retrieving the bit she is interested in.

A PIR scheme with q servers, query size t, answer size l, and probability of correct retrieval $1/2 + \epsilon$ gives a $(q, q, \epsilon/2)$-smooth encoding $C : \{0,1\} \to (\{0,1\}^l)^m$ where $m = O(q2^t/\epsilon)$.


Locally Decodable Codes
Smooth Codes
Batch Codes
Private Information Retrieval


Locally Decodable Codes
Smooth Codes
Batch Codes
Private Information Retrieval

Motivation

Suppose we want to store the string

$$\overbrace{10110100110101}^{14 bits}$$

in devices that can contain only 7 bits.

Option 1.

Use two devices,

$$E(\overbrace{1011010}^{L}\overbrace{0110101}^{R}) = (\overbrace{1011010}^{L}, \overbrace{0110101}^{R})$$

To retrieve 2 bits from the string we may need to retrieve 2 bits from the same device.

Option 2.

Use three devices,

$$E(\overbrace{1011010}^{L}\overbrace{0110101}^{R}) = (\overbrace{1011010}^{L}, \overbrace{0110101}^{R}, \overbrace{1101111}^{L\oplus R})$$

We can retrieve 2 bits from the string reading 1 bit from each device:

$$D_{3,11}(10\mathbf{1}1010\,,\, 011\mathbf{0}0101\,,\, 1101111) = (1,0)$$

$$D_{3,4}(10\mathbf{1}1010\,,\, 011\mathbf{0}0101\,,\, 110\mathbf{1}111) = (1,1)$$

$$D_{10,11}(10\mathbf{1}1010\,,\, 011\mathbf{0}0101\,,\, 11\mathbf{0}1111) = (1,0)$$

Definition (Ishai, Kushilevitz, Ostrovsky, Sahai, 2004)

A (n, N, k, m) batch code over Σ is an encoding function

$$E : \Sigma^n \to (\Sigma^*)^m$$

together with a decoding function (algorithm)

$$D : (\Sigma^*)^m \times \{1, \ldots, n\}^k \to \Sigma^k$$

such that

- $D(E(x), i_1, \ldots, i_k) = (x_{i_1}, \ldots, x_{i_k})$ for all $x \in \Sigma^n$.
- D reads at most 1 symbol from each of the m components of $E(x)$.

The m components of $E(x)$ are called *buckets*.

Definition (Ishai, Kushilevitz, Ostrovsky, Sahai, 2004)

A (n, N, k, m) batch code over Σ is an encoding function

$$E : \Sigma^n \to (\Sigma^*)^m$$

together with a decoding function (algorithm)

$$D : (\Sigma^*)^m \times \{1, \ldots, n\}^k \to \Sigma^k$$

such that

- $D(E(x), i_1, \ldots, i_k) = (x_{i_1}, \ldots, x_{i_k})$ for all $x \in \Sigma^n$.
- D reads at most 1 symbol from each of the m components of $E(x)$.

The m components of $E(x)$ are called *buckets*.

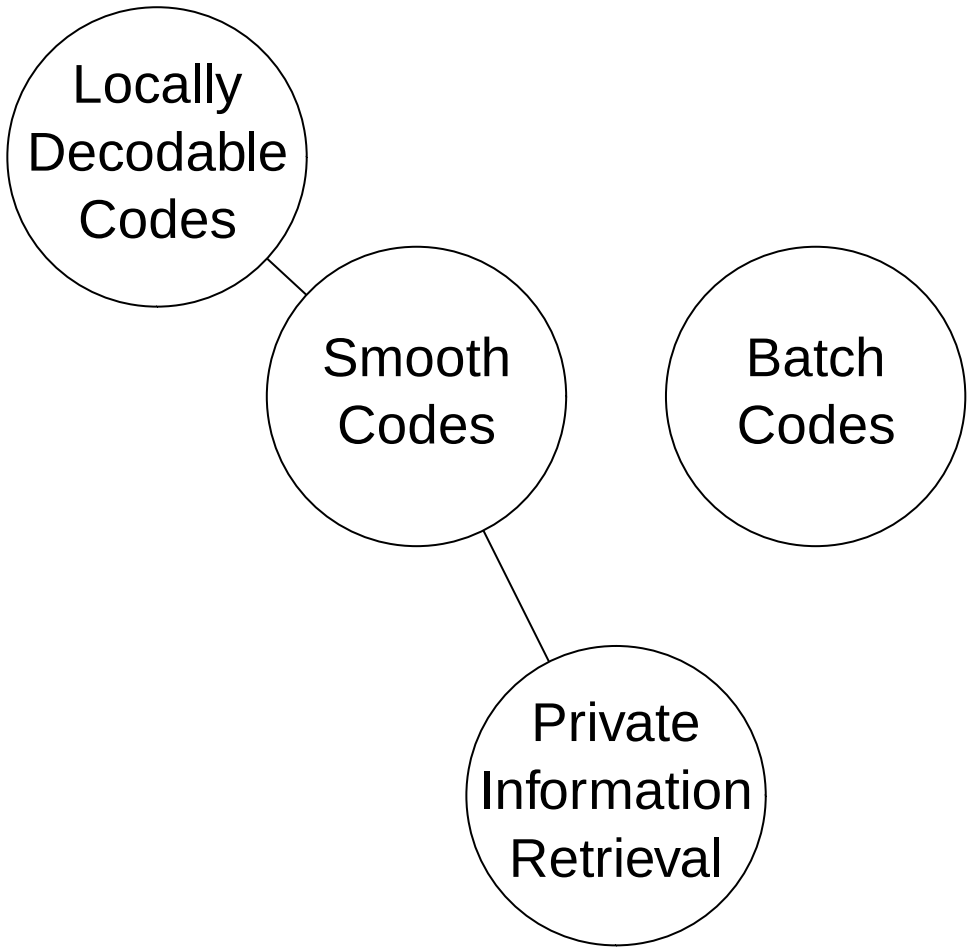
In the previous example $k = 2$, $N = \frac{3}{2}n$, $m = 3$.

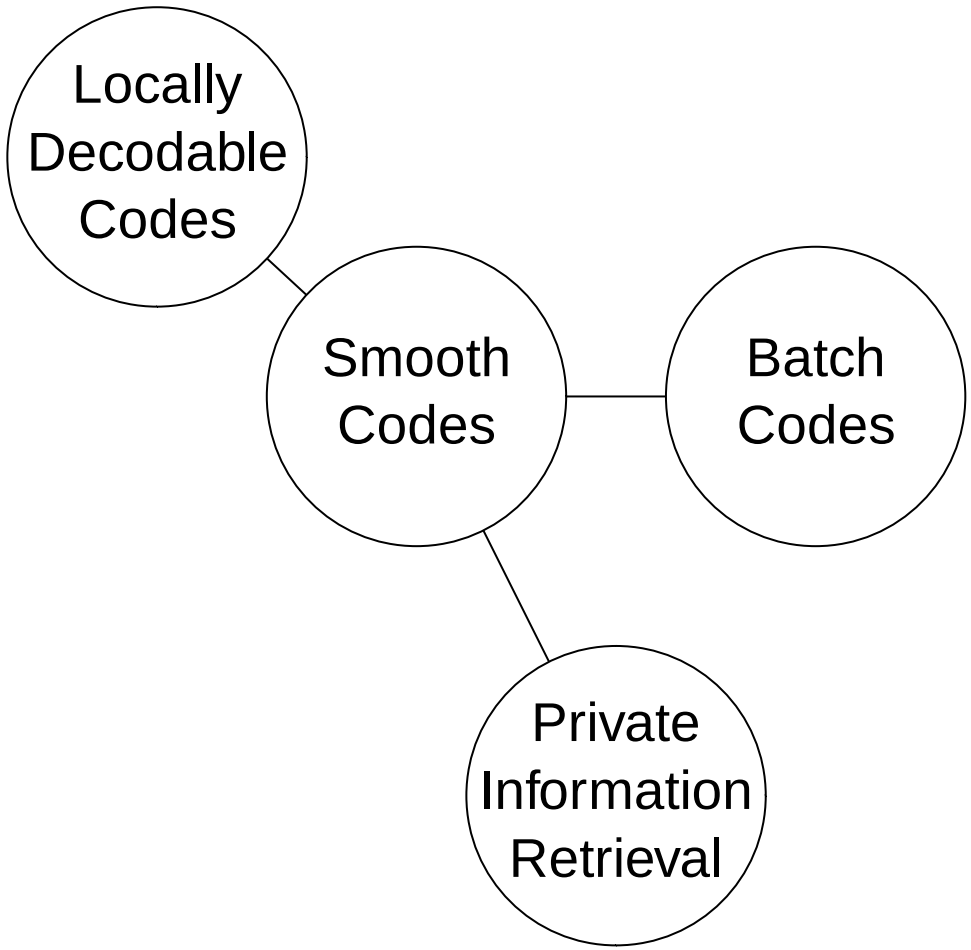
A (n, N, k, m) multiset batch code is a (n, N, k, m) batch code such that for each *k-multi-query* $(i_1, \ldots, i_k)$ the set of m buckets admits a partition of k subsets of buckets such that the i_jth single query can be recovered by reading at most one symbol from each of the buckets in the i_jth subset.

Theorem (Ishai, Kushilevitz, Ostrovsky, Sahai, 2004)

Let $C : \Sigma^n \to \Sigma^m$ be a q-query smooth encoding. Then C describes a $(n, m, \lfloor \frac{m}{q^2} \rfloor, m)$-multiset batch code.

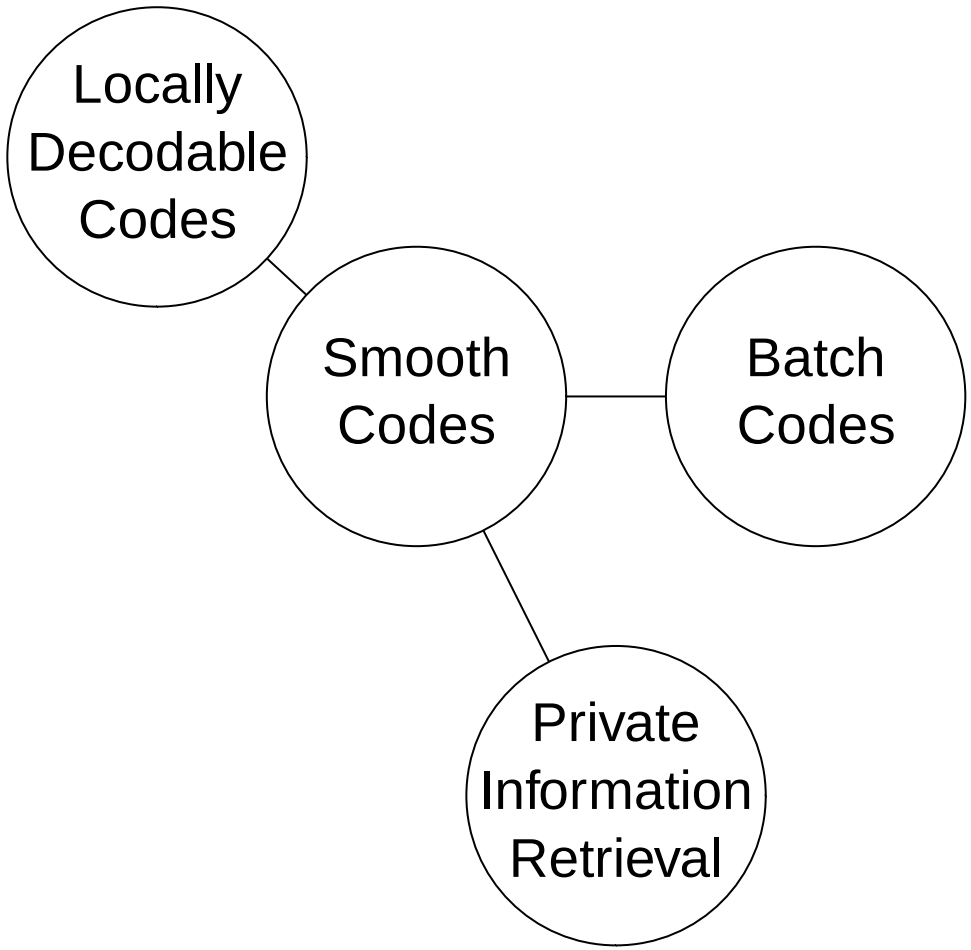
Conversely, the decoding procedure of a (n, m, k, m) multiset batch code gives an *expected* (m/k)-query smooth decoding procedure:
to smoothly decode x_i, run the batch decoder on the multiset $\{i, i, \ldots, i\}$, and pick a random set of buckets from the k disjoint sets allowing to decode x_i.

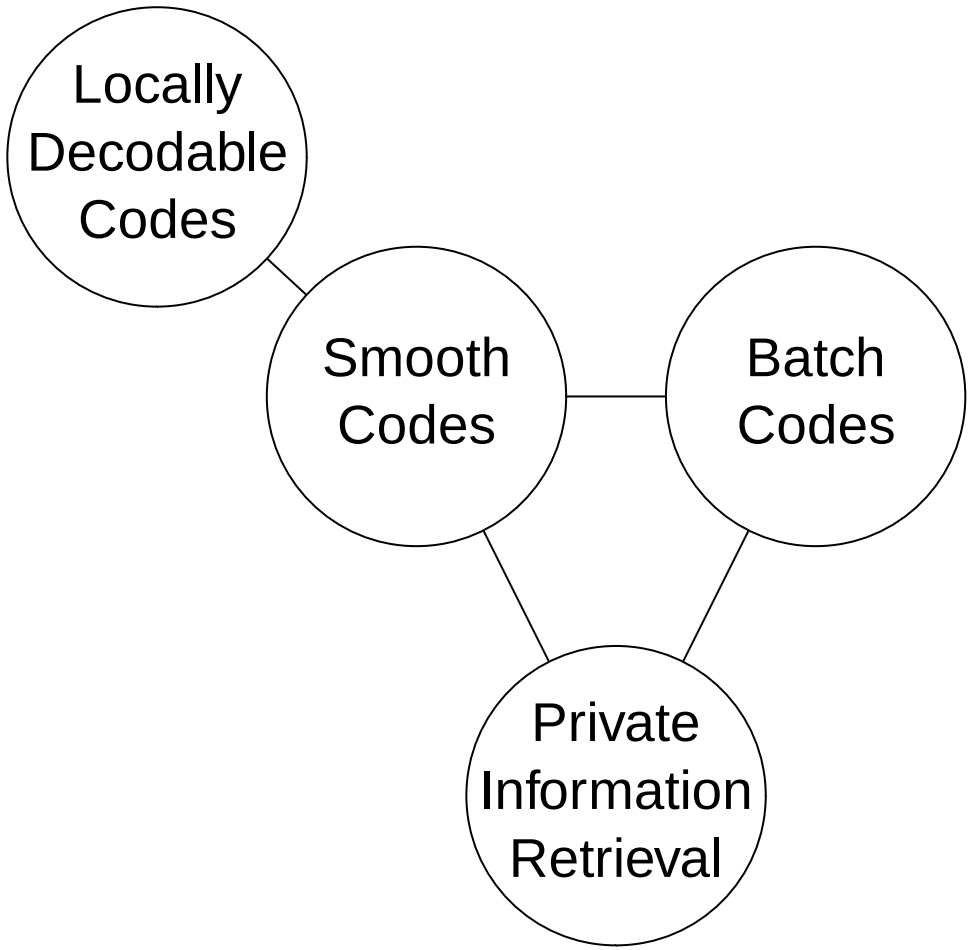

Locally Decodable Codes
Smooth Codes
Batch Codes
Private Information Retrieval


Locally Decodable Codes
Smooth Codes
Batch Codes
Private Information Retrieval

A (n, N, k, m) batch code with bucket sizes N_j, $1 \leqslant k \leqslant m$, provides a reduction from k-query PIR to m invocations of standard PIR on databases of size N_j.

Any nontrivial batch code satisfying $N \ll nk$ implies amortized savings to the time complexity.


Locally Decodable Codes
Smooth Codes
Batch Codes
Private Information Retrieval


Locally Decodable Codes
Smooth Codes
Batch Codes
Private Information Retrieval

Ongoing research topics

- k-PIR
- Peer-to-peer PIR
- Private file sharing system with ECC

What if the database does not cooperate?

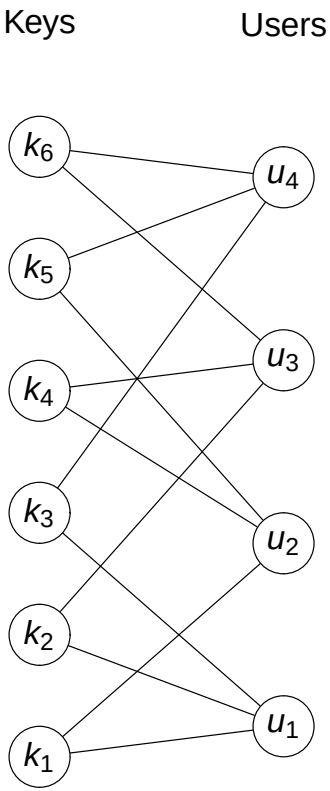
k-PIR: User can mask her target query by ORing it with $k-1$ fake queries.

For maximum privacy,

frequency(target query)=frequency(fake queries).

What if we don’t know the frequencies of all queries?

A configuration is a bipartite graph with constant degree in both sets of vertices and with no cycle of length 4.



A peer-to-peer PIR system can be defined based on the graph.

Private file sharing
needs error correction

n s d f g h j k l c
p h e s c t d x e s
o l p a x r s z g a
i k q l z e a u w p
u j m p d w s l q o
y h n o k k p k m i
t g b i e q j j n u
r f v o h m o b b y
e d y y g n i h u t
w j x t f b u g v g
q a z r d v y f c r

Private file sharing
needs error correction

C	s	d	f	g	h	j	k	l	c
p	**O**	e	s	c	t	d	x	e	s
o	l	**N**	a	x	r	s	z	g	a
i	k	q	**S**	z	e	a	u	w	p
u	j	m	p	**O**	w	s	l	q	o
y	h	n	o	k	**L**	p	k	m	i
t	g	b	i	e	q	**I**	j	n	u
r	f	v	o	h	m	o	**D**	b	y
e	d	y	y	g	n	i	h	**E**	t
w	j	x	t	f	b	u	g	v	**R**
q	a	z	r	d	v	y	f	c	r

Private file sharing
needs error correction

C	s	d	f	g	h	j	k	l	c
p	**O**	e	s	c	t	d	x	e	s
o	l	**N**	a	**C**	**A**	**S**	**T**	**R**	**O**
i	k	q	**S**	z	e	a	u	w	p
u	j	m	p	**O**	w	s	l	q	o
y	h	n	o	k	**L**	p	k	m	i
t	g	b	i	e	q	**I**	j	n	u
r	f	v	o	h	m	o	**D**	b	y
e	d	y	y	g	n	i	h	**E**	t
w	j	x	t	f	b	u	g	v	**R**
q	a	z	r	d	v	y	f	c	r

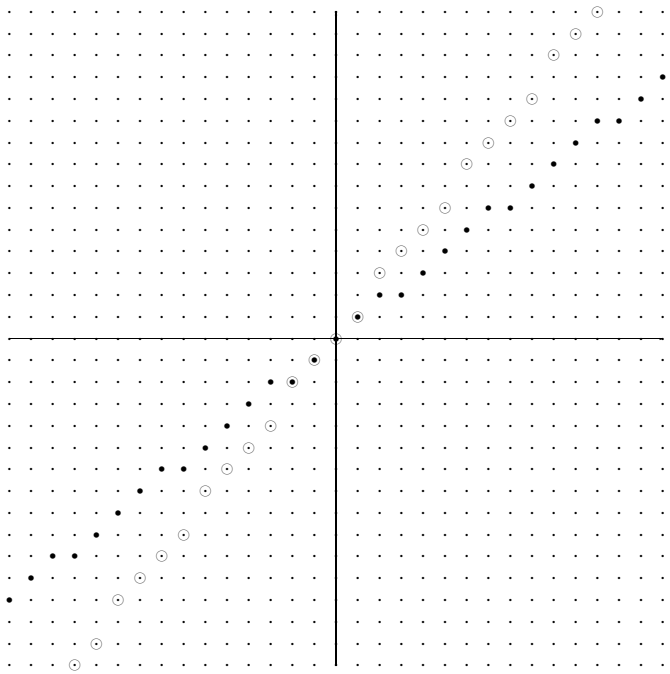
Private file sharing
needs error correction

C	s	d	f	g	h	j	k	l	c
p	**O**	e	s	c	t	d	x	e	s
o	l	**N**	a	**C**	**A**	**S**	**T**	**S**	**O**
i	k	q	**S**	z	e	a	**E**	w	p
u	j	m	p	**O**	w	**L**	l	q	o
y	h	n	o	k	**A**	p	k	m	i
t	g	b	i	**I**	q	**I**	j	n	u
r	f	v	**D**	h	m	o	**D**	b	y
e	d	**R**	y	g	n	i	h	**E**	t
w	**U**	x	t	f	b	u	g	v	**R**
q	a	z	r	d	v	y	f	c	r

Private file sharing needs error correction

Problems!

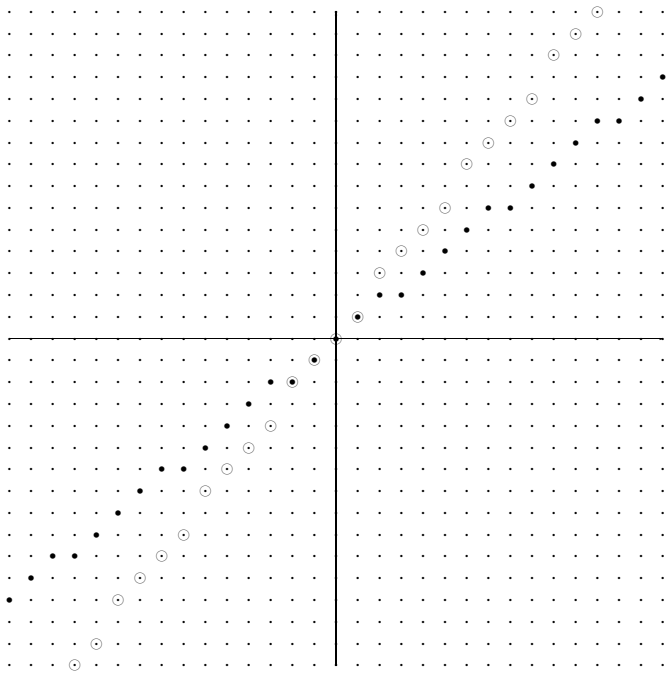
C	s	d	f	g	h	j	k	l	c
p	**O**	e	s	c	t	d	x	e	s
o	l	**N**	a	**C**	**A**	**S**	**T**	**S**	**O**
i	k	q	**S**	z	e	a	**E**	w	p
u	j	m	p	**O**	w	**L**	l	q	o
y	h	n	o	k	**A**	p	k	m	i
t	g	b	i	**I**	q	**I**	j	n	u
r	f	v	**D**	h	m	o	**D**	b	y
e	d	**R**	y	g	n	i	h	**E**	t
w	**U**	x	t	f	b	u	g	v	**R**
q	a	z	r	d	v	y	f	c	r



Private file sharing needs error correction

Problems!

C	s	d	f	g	h	j	k	l	c
p	**O**	e	s	c	t	d	x	e	s
o	l	**N**	a	**C**	**A**	**S**	**T**	**S**	**O**
i	k	q	**S**	z	e	a	**E**	w	p
u	j	m	p	**O**	w	**L**	l	q	o
y	h	n	o	k	**A**	p	k	m	i
t	g	b	i	**I**	q	**I**	j	n	u
r	f	v	**D**	h	m	o	**D**	b	y
e	d	**R**	y	g	n	i	h	**E**	t
w	**U**	x	t	f	b	u	g	v	**R**
q	a	z	r	d	v	y	f	c	r



Alternative: combinatorial structures