User privacy

Vicenç Torra

February, 2018

SAIL + PICS, School of Informatics, University of Skövde, Sweden

Outline

1. User privacy

User privacy

Data Privacy

Classification 1: On whose privacy is being sought

- Respondent privacy
- Owner privacy
- User privacy

Data Privacy

User privacy

- Protecting the identity of the user
- Protecting the data generated by the activity of the user

Data Privacy

User privacy

- Protecting the identity of the user
- Protecting the data generated by the activity of the user

Tools for anonymous communications belong to user privacy

Data Privacy

User privacy

- Protecting the identity of the user
- Protecting the data generated by the activity of the user

Tools for anonymous communications belong to user privacy

Other examples with users querying databases

Data Privacy

User privacy in database search

- Protecting the identity of the user
  - Protect who is making a query

Data Privacy

User privacy in database search

- Protecting the identity of the user
  - Protect who is making a query

    $\rightarrow$ Anonymous database search
- Protecting the data generated by the user

Data Privacy

User privacy in database search

- Protecting the identity of the user
  - Protect who is making a query
    $\rightarrow$ Anonymous database search
- Protecting the data generated by the user
  - Protect the query of the user

Data Privacy

User privacy in database search

- Protecting the identity of the user
  - Protect who is making a query
    $\rightarrow$ Anonymous database search
- Protecting the data generated by the user
  - Protect the query of the user
    $\rightarrow$ Private Information Retrieval (PIR)

Data Privacy

User privacy

- Private Information Retrieval (PIR)
- Anonymous database search

Data Privacy

User privacy

- Private Information Retrieval (PIR)
  - How a user should retrieve an element from a DB or a search engine, without the system or the server being able to deduce which element is the object of the user's interest.

Data Privacy

User privacy

- Private Information Retrieval (PIR)
  - (Information Theoretic) Private Information Retrieval (PIR)
  - Computational PIR (cPIR)
  - Trusted-hardware PIR
  - Other approaches
    - ⋆ Goopir
    - ⋆ TrackMeNot

Data Privacy

User privacy

- (Information Theoretic) Private Information Retrieval (PIR)
  - Information theoretic: cannot be broken with unlimited computing power

Data Privacy

User privacy

- (Information Theoretic) Private Information Retrieval (PIR)
  - Information theoretic: cannot be broken with unlimited computing power
  - Every (information theoretic) PIR scheme with a single-database (with n bits) requires $\Omega(n)$ bits of communication.

Data Privacy

User privacy

- (Information Theoretic) Private Information Retrieval (PIR)
  - Information theoretic: cannot be broken with unlimited computing power
  - Every (information theoretic) PIR scheme with a single-database (with n bits) requires $\Omega(n)$ bits of communication.
  - It can be proven (Chor et al. 1998) that if a user wants to keep its privacy (in the information theoretic sense), then essentially the only thing he can do is to ask for a copy of the whole database.

Data Privacy

User privacy

- (Information Theoretic) PIR:
  - Communication complexity is reduced: sublinear in n by assuming that the data is replicated.

Data Privacy

User privacy

- (Information Theoretic) PIR:
  - Communication complexity is reduced: sublinear in n by assuming that the data is replicated.
    - $\star$ k copies of the database are considered
    - $\star$ DB copies do not collaborate

Data Privacy

User privacy

- (Information Theoretic) PIR:
  - Communication complexity is reduced: sublinear in n by assuming that the data is replicated.
    - $\star$ k copies of the database are considered
    - $\star$ DB copies do not collaborate
  - **Example.** Scheme in (Chor et al., 1999) with communication complexity $O(n^{1/3})$ for $k = 2$

Data Privacy

User privacy

- (Information Theoretic) PIR: k copies of the database (not being intercommunicated)
  - Problem.
    - ⋆ Database. A binary string $x = x_1 \cdots x_n$ of length n
      (Identical copies of this string are stored in $k \geq 2$ servers)
    - ⋆ User. Given index i, is interested in obtaining the value of bit x_i
    - ⋆ *Solution:* The user queries each of the servers and gets replies from which the desired bit x_i can be computed.
      The server does not gain any information about i from the query.

Data Privacy

Definition of the problem. (Information Theoretic) PIR (I)

- Input
  - $i \in [n]$ where $[n] = \{1, \ldots, n\}$
  - r random input of length ℓ_{rnd}
- Overview of the process
  - k queries $Q_1(i, r), \ldots, Q_k(i, r)$ of length ℓ_q each
  - Servers respond according to strategies $A_1, \ldots, A_k$ with replies of length ℓ_a according to the content of the DB x
  - The user reconstructs the desired bit x_i from the k replies, together with i and r

Data Privacy

Definition of the problem. (Information Theoretic) PIR (I)

- Formalization
  - A k-server PIR scheme for database length n consists of
    - $\star$ k query functions $Q_1, \ldots, Q_k : [n] \times \{0,1\}^{\ell_{rnd}} \rightarrow \{0,1\}^{l_q}$
    - $\star$ k answer functions, $A_1, \ldots, A_k : \{0,1\}^n \times \{0,1\}^{l_q} \rightarrow \{0,1\}^{l_a}$
    - $\star$ a reconstruction function $R : [n] \times \{0,1\}^{l_{rnd}} \times (\{0,1\}^{l_a})^k \rightarrow \{0,1\}$
  - These functions should satisfy

Data Privacy

Definition of the problem. (Information Theoretic) PIR (I)

- Formalization
  - A k-server PIR scheme for database length n consists of
    - ⋆ k query functions $Q_1, \ldots, Q_k : [n] \times \{0,1\}^{\ell_{rnd}} \rightarrow \{0,1\}^{l_q}$
    - ⋆ k answer functions, $A_1, \ldots, A_k : \{0,1\}^n \times \{0,1\}^{l_q} \rightarrow \{0,1\}^{l_a}$
    - ⋆ a reconstruction function $R : [n] \times \{0,1\}^{l_{rnd}} \times (\{0,1\}^{l_a})^k \rightarrow \{0,1\}$
  - These functions should satisfy
    - ⋆ Correctness. For every $x \in \{0,1\}^n$, $i \in [n]$, and $r \in \{0,1\}^{\ell_{rnd}}$

$$R(i, r, A_1(x, Q_1(i,r)), \ldots, A_k(x, Q_k(i,r))) = x_i$$

Data Privacy

Definition of the problem. (Information Theoretic) PIR (I)

- Formalization
  - A k-server PIR scheme for database length n consists of
    - ⋆ k query functions $Q_1, \ldots, Q_k : [n] \times \{0,1\}^{\ell_{rnd}} \rightarrow \{0,1\}^{l_q}$
    - ⋆ k answer functions, $A_1, \ldots, A_k : \{0,1\}^n \times \{0,1\}^{l_q} \rightarrow \{0,1\}^{l_a}$
    - ⋆ a reconstruction function $R : [n] \times \{0,1\}^{\ell_{rnd}} \times (\{0,1\}^{l_a})^k \rightarrow \{0,1\}$
  - These functions should satisfy
    - ⋆ Correctness. For every $x \in \{0,1\}^n$, $i \in [n]$, and $r \in \{0,1\}^{\ell_{rnd}}$
      $$R(i, r, A_1(x, Q_1(i,r)), \ldots, A_k(x, Q_k(i,r))) = x_i$$
    - ⋆ Privacy. For every $i, j \in [n]$, $s \in [k]$, and $q \in \{0,1\}^{l_q}$
      $$Pr(Q_s(i,r) = q) = Pr(Q_s(j,r) = q)$$
      where the probabilities are taken over uniformy chosen $r \in \{0,1\}^{\ell_{rnd}}$

Data Privacy

User privacy

- (Information Theoretic) PIR: k copies of the database (not being intercommunicated)
  - Variations.
    - ⋆ Protocols can be defined to coalitions of up to $t < k$ servers

Data Privacy

User privacy

- Computational PIR (cPIR): privacy against one single database
  - The server has limited computational capacity
    - ⋆ The computations the server has to perform in order to gather enough information on the searches of a user to vulnerate her privacy, exceeds the capacity of the server.

Data Privacy

User privacy

- Computational PIR (cPIR): privacy against one single database
  - First approaches:
    - (Chor, Gilboa, 1997) For every $0 < c < 1$ there is a cPIR scheme for $k = 2$ DB with communication complexity $O(n^c)$.
    - (Kushilevitz, Ostrovsky, 1997) For every $c > 0$ there exists a single-database cPIR scheme with communication complexity $O(n^c)$, assuming the hardness of deciding quadratic residuosity[1]. Linear time for the DB with respect to the number of rows.

    $\rightarrow$ They present a basic scheme and a recursive scheme

[1] Given (x, N) where N is a composite number, it is difficult to determine whether x is a quadratic residue modulo N (i.e., $x = y^2 \ mod\ N$ for a certain y).

Data Privacy

User privacy

- Trusted-hardware Private Information Retrieval (hardware-based Private Information Retrieval)
  - PIR protocols based on the assumption of a trusted hardware

Data Privacy

User privacy

- Other systems
  - Goopir: A user masks the query with $k-1$ fake queries (example: change w_1 by $w_1 or w_2 or \ldots or w_k$) and submit the query to the search engine
    - ⋆ It assumes that frequencies of keywords and phrases that appear in a query are known in advance.
      $\rightarrow$ the frequencies of the target and the fake queries should be similar
      so that the uncertainty of the search engine about the real target query is maximum
      $\rightarrow$ maximum privacy

Data Privacy

User privacy

- Other systems
  - TrackMeNot: A plugin for Firefox that periodically issues search queries
    $\rightarrow$ it hides the users actual search trails in a cloud of *ghost* queries.
    - ⋆ Generalization of its use: overhead of ghost queries
    - ⋆ Automatic ghost queries might be distinguishable and provide clues

Data Privacy

User privacy

- Private Information Retrieval (PIR)
- Anonymous database search

Data Privacy

User privacy

- Anonymous database search
  - How a user should retrieve an element from a database or a search engine without the system or the server being able to deduce who the retrieving user is.
    $\rightarrow$ It does not hide the content of the query, but obstructs the possibilities for the database of profiling users.

Data Privacy

User privacy

- P2P UPIR: Peer-to-peer User-Private Information Retrieval
  - Users submit queries on behalf of other users
  - The way in which users share communication spaces (memory sectors and cryptographic keys) is defined using combinatorial configurations
  - P2P UPIR offers privacy versus peer users

Data Privacy

P2P UPIR: Peer-to-peer User-Private Information Retrieval

- Communities of users and communication space: case 1
  - one memory sector and one cryptographic key
    - ⋆ all write and read
    - ⋆ the DB cannot know who is asking what: no profiling (except for the group)

      → but, no privacy between users

      The user does not know who made the query, but all queries are known

Data Privacy

P2P UPIR: Peer-to-peer User-Private Information Retrieval

- Communities of users and communication space: case 2
  - each user shares a different communication space with every other user
    - ⋆ every user only reads requests from “neighbours”
      - → The user knows who requested a query, and its content
      - → Not all the queries are known

Data Privacy

P2P UPIR: Peer-to-peer User-Private Information Retrieval

- Communities of users and communication space: case 3
  - different communication spaces for different users
    - ⋆ n_c communication spaces

      with a memory sector and a cryptographic key
    - ⋆ n_u a set of users

      all of them having access to a subset of d_u communication spaces

      so that every communication space is shared by d_c users

      and every pair of users share at most one communication space

$case 1$	$case 2$
$n_c = 1$ (one space)	$n_c = \frac{n_u(n_u-1)}{2}$ (one space for each pair)
$d_u = 1$ (one space per user)	$d_u = n_u - 1$ (for each user, one space for each other user)
$d_c = n_u$ (the only space is shared by all users)	$d_c = 2$ (each space: only two users)