T-79.514 Special Course on Cryptology

Private Information Retrieval

Vesa Vaskelainen
Helsinki University of Technology
vvaskela@cc.hut.fi

Overview of the Lecture

- Private Information Retrieval (PIR)

  ⋆ Allow a user to retrieve information from a database while maintaining his query private

- Symmetrically Private Information Retrieval (SPIR)

  ⋆ Quarantees also the privacy of the data, as well as of the user

- Very Short Introduction to Quantum Mechanics

  ⋆ Formalism used in quantum computing

- Quantum SPIR scheme on top of the classical PIR scheme

Background

- Data privacy is a natural and crucial requirement in many settings. For example, consider a commercial database which sells information, such as stock information, to users, charging by the amount of data that the user retrieved. Here, both user privacy and database privacy are essential.

- Y. Gertner et al. *Protecting Data Privacy in Private Information Retrieval Schemes*. Journal of Computer and Systems Sciences, 60(3):592–629, 2000. Earlier version in STOC 98.

- I. Kerenidis, R. de Wolf. *Quantum Symmerically-Private Information Retrieval*. arXiv:quant-ph/0307076, 2003.

Definitions

- Database $\mathcal{DB}$ is a binary string $x = x_1 \ldots x_n$ of length n, identical copies of this string are stored by $k \geq 2$ servers

- By $[l]$ is denoted the set $\{1, 2, \ldots, l\}$. For any sets $S, S' \subseteq [l]$, we let $S \oplus S'$ denote the symmetric difference between S and S' (i.e., $S \oplus S' = (S \backslash S') \cup (S' \backslash S)$), and χ_S denote the characteristic vector of S: an l-bit binary string whose j-th bit is equal to 1 iff $j \in S$.

- $\{0, 1\}^n$ is the set of strings of length n with each letter being either zero or one.

- “PIR and SPIR scheme” refer to 1-round information theoretically private schemes

- Complexity is measured in terms of communication

- *User privacy* requirement: under any two indices i, i', the communication seen by any single database is identically distributed

- The *data privacy* condition of SPIR schemes requires for any user interacting with the honest databases $\mathcal{DB}_1, \ldots, \mathcal{DB}_k$ there exists an index i s.t. for every data strings x, x' satisfying $x_i = x'_i$ the distribution of communication is independent of the data strings x and x'.

Basic Cube Scheme

$k = 2^d$ databases, the size of $n = l^d$, where $d, l \in \mathbb{Z}_+$. The index set $[n]$, is identified with the d-dimensional cube $[l]^d$. Each index $i \in [n]$, is identified with a d-tuple $(i_1, \ldots, i_d)$. A d-dimensional subcube $S_1 \times \cdots \times S_d \subseteq [l]^d$, where each $S_i \subseteq [l]$.

QUERIES: The user picks a random $(S_1^0, \ldots, S_d^0)$, where $S_1^0, \ldots, S_d^0 \subseteq [l]$. Let $S_m^1 = S_m^0 \oplus i_m$ $(1 \leq m \leq d)$. For each $\sigma = \sigma_1\sigma_2 \ldots \sigma_d \in \{0, 1\}^d$, the user sends to $\mathcal{DB}_\sigma$ the subcube $C_\sigma = (S_1^{\sigma_1}, \ldots, S_d^{\sigma_d})$, where each $S_m^{\sigma_m}$ is presented by its characteristic l-bit string.

ANSWERS: Each $\mathcal{DB}_\sigma$, $\sigma \in \{0, 1\}^d$, computes XOR of the bits in the subcube C_σ, and sends the resultant bit b_σ to the user.

RECONSTRUCTION: The user computes $x_i = \bigoplus_{\sigma \in \{0,1\}^d} b_\sigma$.

PIR Scheme $\mathcal{B}_2$ (2-database covering-codes scheme)

$l = n^{1/3}$, $i = (i_1, i_2, i_3)$, $\mathcal{DB}_{000}$ and $\mathcal{DB}_{111}$ emulates the 4 databases $\mathcal{DB}_\sigma$, $\sigma \in \{0,1\}^3$, s.t. Hamming distance of σ from its index is at most 1.

QUERIES: The user sends $C_{000} = (S_1^0, S_2^0, S_3^0)$ to $\mathcal{DB}_{000}$ and $C_{111} = (S_1^1, S_2^1, S_3^1)$ to $\mathcal{DB}_{111}$.

ANSWERS: $\mathcal{DB}_{000,111}$ replies with single bits $b_{000,111}$ along with 3 l-bit long strings, i.e. $\mathcal{DB}_{000}$ emulates $\mathcal{DB}_{100}$ by computing $\bigoplus(S_1^0 \oplus i_1, S_2^0, S_3^0)$ for every $i_1 \in [l]$.

RECONSTRUCTION: In the l-bit long strings, the index of the required answer bit b_σ is i_1 (for $\sigma = 100, 011$), i_2 ($\sigma = 010, 101$), or i_3 ($\sigma = 001, 110$). The user computes $x_i = \bigoplus_{\sigma \in \{0,1\}^3} b_\sigma$.

Correctness and Complexity

- The correctness of the basic cube scheme follows from the fact that every bit in x except x_i appears in an even number of subcubes C_σ, $\sigma \in \{0,1\}^d$, and x_i appears in exactly one such subcube.

- For the basic cube scheme communication complexity is $k \cdot (d \cdot l + 1) = 2^d \cdot (d \cdot \sqrt[d]{n} + 1) = \mathcal{O}(n^{1/d})$

- $\mathcal{B}_2$ has total communcation complexity $2(6\sqrt[3]{n}+1) = \mathcal{O}(n^{1/3})$. Note that it is too expensive to let $\mathcal{DB}_{000}$ emulate $\mathcal{DB}_{011}$ as this will require considering all $(\sqrt[3]{n})^2$ possibilities for (S_2^1, S_3^1).

Conditional Disclosure of Secrets

- The “condition” h: $\{0,1\}^n \to \{0,1\}$ for some n; an external party Carol holds $y \in \{0,1\}^n$, which is also partitioned between the $P_1, \ldots, P_k$ players which have access to a shared random string (hidden from Carol). A secret input s is known to at least one of the players. Based on its share of y and on the shared randomness, each P_j simultaneously sends a message to Carol, s.t. (1) if $h(y) = 1$, then Carol is able to reconstruct the secret s; and (2) if $h(y) = 0$, then Carol obtains no information about s.

- **Claim 1.** *Suppose h: $\{0,1\}^n$ has a Boolean formula of size $S(n)$, and let s denote a secret bit known to at least one player. Then there exist a protocol for disclosing s subject to the condition h, whose total communication complexity is $S(n) + 1$.*

Private Simultaneous Messages (PSM)

- Each player $P_1, \ldots, P_k$ is holding a private input string y_j. All players have access to a shared random input, which is unknown to Carol. Based on y_j and the shared random input, each player P_j simultaneously sends a single message to Carol. From the messages she received, Carol should be able to compute some predetermined function $f(y_1, \ldots, y_k)$, but should obtain no additional information on the input other than what follows from the value of f.

- **Example 1.** In the basic cube scheme data privacy can be maintained (respect to an honest user) if instead of sending original answer b_σ, each $\mathcal{DB}_\sigma$ sends a masked answer $b_\sigma \oplus r_\sigma$, where $r = r_{0\ldots00} r_{0\ldots01} \cdots r_{1\ldots11}$ are randomly chosen from the k-tuples whose bits XOR to 0.

Honest-User-SPIR Schemes $\mathcal{B}'_2$ and $\mathcal{B}'_k$

- The reconstruction function of $\mathcal{B}_2$ may be viewed as a two-stage procedure: (1) the user selects a single bit from each of 8 answer strings, depending only on the index i; and (2) the user exclusive-ors the 8 bits it has selected to obtain x_i.

- The user independently shares χ_{i_m}, $m = 1,2,3$, among the two databases. ($r_m^0 \oplus r_m^1 = \chi_{i_m}$)

- Each bit of a_σ is an input to a PSM protocol computing the XOR of 8 answer bits. Let w_σ denote the string where each bit from a_σ is replaced by its corresponding PSM message bit.

- For every $\sigma \in \{0,1\}^3$ and $1 \leq j \leq |w_\sigma|$, the database use their shared randomness to disclose to the user the j-th bit of w_σ, $(w_\sigma)_j$, subject to an appropriate condition $(r_m^0)_j \oplus (r_m^1)_j = 1$.

- The user reconstructs the eight PSM message bits corresponding to the index i (using the reconstruction function of the conditional disclosure protocol), and computes their exclusive-or to obtain x_i.

- Based on the **Claim 1.** it can be shown that the communication complexity of the $\mathcal{B}_2'$ is $\mathcal{O}(n^{1/3})$. Generalization gives,

  Theorem 1. *For every constant $k \geq 2$ there exist a k-database honest-user-SPIR scheme, $\mathcal{B}_k'$, of communication complexity $\mathcal{O}(n^{1/(2k-1)})$.*

Cube Schemes $\mathcal{B}_2''$ and $\mathcal{B}_k''$

- The user can cheat in two ways in the previous honest-user-SPIR scheme: sharing the all-ones vector instead of χ_{i_m}, and by sending invalid queries invalid queries in the original PIR scheme. (may obtain $\mathcal{O}(n^{1/3})$ physical data bits)

- The databases share a random bit s. The bit s is disclosed to the user subject to the condition $\bigwedge_{m=1}^{3}(S_m^0 \oplus S_m^1 = \{r_m^0 \oplus r_m^1\})$ which validates the user's queries.

- The honest user can reconstruct s and the 8 bits corresponding to index i and compute their exclusive-or to obtain x_i. The user can only learn $(s \oplus b_{000} \oplus b_{111} \oplus b)$, where $b = \bigoplus_{\sigma \neq 000,111} b_\sigma$.

- The user's queries can be verified by a Boolean formula of size $\mathcal{O}(l \log l)$. For disclosing PSM message strings w_σ one needs a Boolean formula of size $\mathcal{O}(\log l)$. From these it follows that the scheme $\mathcal{B}_2''$ has communication complexity $\mathcal{O}(\log n \cdot n^{1/3})$.

- The previous is generalized by the following theorem.

  Theorem 2. *For every constant $k \geq 2$ there exist a k-database SPIR scheme, $\mathcal{B}_k''$, of communication complexity $\mathcal{O}(\log n \cdot n^{1/(2k-1)})$.*

Very Short Introduction to Quantum Mechanics

- The standard quantum mechanical notation for a vector in a complex vector space is $|\psi\rangle$
- The quantum analog of a bit is *qubit* which is two- state system where the two possible states are called $|0\rangle$ and $|1\rangle$.
- The most essential property of them is the possibility of superposition. The general state is, $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ where $|\alpha|^2 + |\beta|^2 = 1$.
- The elements of $V \otimes W$ are linear combinations of 'tensor products' $|v\rangle \otimes |w\rangle$ of elements $|v\rangle$ of V and $|w\rangle$ of W.

QSPIR Scheme

The user picks a random string r, and depending on index i and r, picks k queries $q_1, \ldots, q_k \in \{0,1\}^t$. In addition, he picks k random strings $r_1, \ldots, r_k \in \{0,1\}^a$. The user also holds strings $b_1, \ldots, b_k \in \{0,1\}^a$ which are determined by i and r in a way that

$$\sum_{j=1}^{k} a_j \cdot b_j = x_i \qquad (\mathrm{mod}\ 2).$$

The user defines $r'_j = r_j - b_j$ and set up the following $(1 + k(t+a))$-qubit state

$$\frac{1}{\sqrt{2}}|0\rangle|q_1, r_1\rangle \ldots |q_k, r_k\rangle + \frac{1}{\sqrt{2}}|q_1, r'_1\rangle \ldots |q_k, r'_k\rangle.$$

The jth server performs the following unitary mapping,

$$|q_j, r\rangle \rightarrow (-1)^{a_j \cdot r}|q_j, r\rangle.$$

The servers then send all the qubits they have back to the user.

$$\frac{1}{\sqrt{2}}(-1)^{a_1 \cdot r_1}|q_1, r_1\rangle \dots (-1)^{a_k \cdot r_k}|q_k, r_k\rangle$$

$$+\frac{1}{\sqrt{2}}(-1)^{a_1 \cdot r'_1}|q_1, r'_1\rangle \dots (-1)^{a_k \cdot r'_k}|q_k, r'_k\rangle.$$

The common factor $(-1)^{\sum_j a_j \cdot r_j}$ can be ignored. Thus previous equals to,

$$\frac{1}{\sqrt{2}}|0\rangle|q_1, r_1\rangle \dots |q_k, r_k\rangle + \frac{1}{\sqrt{2}}|1\rangle(-1)^{\sum_{j=1}^{k} a_j \cdot b_j}|q_1, r'_1\rangle \dots |q_k, r'_k\rangle =$$

$$\frac{1}{\sqrt{2}}|0\rangle|q_1, r_1\rangle \dots |q_k, r_k\rangle + \frac{1}{\sqrt{2}}|1\rangle(-1)^{x_i}|q_1, r'_1\rangle \dots |q_k, r'_k\rangle.$$

The user can get $|x_i\rangle$ from this by using Hadamard transform operator

$$H \equiv \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Conclusions

- Clearly, PIR can be realized by making the server send the whole database to user, better protocols exist if the database is replicated among some $k \geq 2$ different servers, who cannot communicate.

- Classical SPIR schemes requires the shared randomness between servers.

- The honest-user quantum SPIR schemes exist even in the case where the servers do not share any randomness.