Security Analysis of OXT

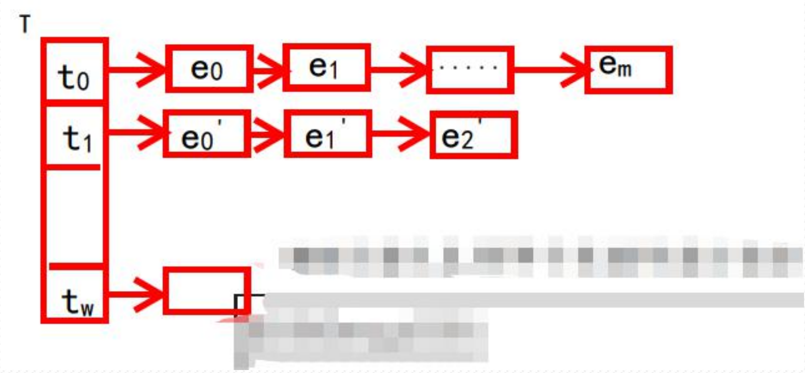
《Highly-Scalable Searchable Symmetric Encryption with Support forBoolean Queries》

Zhou Jiancong
2022.3.21

提纲

- 回顾OXT方案
- 证明思路
- 预备知识
- 安全性分析
- 初步Idea

回顾OXT方案



SKS

【优势】将每个关键词以密态形式存储在TSet中，具备基础的SSE能力
【不足】只支持但关键词检索，**不支持连词检索**

BXT

【优势】对w1和其他连词进行差异处理，对w1在TSet中检索，对其余关键词在XSet中进行伪随机函数（哈希）匹配，**提升了检索的效率**
【不足】不安全的信道中缺少有效的密钥传输机制，导致**ind以明文方式传输造成泄露**

OXT

【优势】通过DH非对称结构解决了ind结果以明文传输的缺陷，**兼顾效率的基础上提升了安全性**
【不足】方案**效率下降**，并且完美解决**否定逻辑问题**。

安全性分析-OXT的安全性证明

- 证明思路：

最终目的是证明boolean查询在自适应方案下的安全性，首先通过非自适应性方案下关键词两两连接特例的安全性进行证明。特例证明基本足以覆盖最终证明的难点。

OXT是语义安全的，能够抵抗2个连词的非适应性攻击

Theorem 5 *Let $\mathcal{L}_{\text{oxt}}$ be as defined above, and suppose that the T-set implementation Σ from Section 6. Then SSE scheme OXT is $\mathcal{L}_{\text{oxt}}$-semantically-secure against non-adaptive attacks where are all queries are 2-conjunctions, assuming that the DDH assumption holds in G, that F and F_p are secure PRFs, that (Enc, Dec) is an IND-CPA secure symmetric encryption scheme, and the conditions from Theorem 7 hold.*

PRF Security. Let X and Y be sets, and let $F : \{0,1\}^\lambda \times X \to Y$ be a function. We say that F is a *pseudorandom function (PRF)* if for all efficient adversaries A, $\mathbf{Adv}^{\text{prf}}_{F,A}(\lambda)$ is negligible, where

$$\mathbf{Adv}^{\text{prf}}_{F,A}(\lambda) = \Pr[A^{F(K,\cdot)}(1^\lambda) = 1] - \Pr[A^{f(\cdot)}(1^k) = 1]$$

where the probability is over the randomness of A, $K \xleftarrow{\$} \{0,1\}^\lambda$, and $f \xleftarrow{\$} \mathsf{Fun}(X, Y)$.

前置条件1：伪随机函数(PRF)存在

Lemma 3 *Suppose the DDH assumption holds for in G. Then, for any integers α, β (polynomial in λ) any efficient adversary A, we have*

$$\Pr[A(g, g^{\mathbf{a}}, g^{\mathbf{b}}, g^{\mathbf{a}\mathbf{b}^T}) = 1] - \Pr[A(g, g^{\mathbf{a}}, g^{\mathbf{b}}, \mathbf{M}) = 1] \le \text{neg}(\lambda),$$

where $\mathbf{a}$ is uniform over $(Z_p^*)^\alpha$, $\mathbf{b}$ is uniform over $(Z_p^*)^\beta$, and $\mathbf{M}$ is uniform over $G^{\alpha\times\beta}$.

前置条件2：DDH假设成立

Lemma 4 *For every adversary A there exists adversaries B and B' which run in essentially the same time as A, such that*

$$\Pr[\mathbf{Cor}^{\mathsf{OXT}}_A(\lambda) = 1] \le 2 \cdot \mathbf{Adv}^{\text{prf}}_{F_p,B}(\lambda) + \mathbf{AdvCor}^{\Pi}_{B'}(\lambda) + N^2/(p-1) + N/p,$$

where $N = \Sigma_{i=1}^d |\mathsf{W}_i|$ is the total number of appearances of keywords in all documents, p is the order of the group G, and Π is the T-set implementation.

前置条件3：引理4得证（证明过程见附录）

预备知识

- 群（Group）

  群是集合 G 与一运算 * 的结合体，且满足以下条件：

  (1)群中有一个**单位元**（也称幺元），集合G中存在元素e，对 G 中任意元素 a,满足 a *e= e* a = a

  (2)群的元素有有**逆元**, 对 G 中任意元素 a，存在b满足 a * b = e 和 b * a = e

  (3)运算满足**结合律**，对 G 中任意元素 a、b、c，满足a * (b * c) = (a * b) * c

  (4)群对运算是**封闭的**, 对 G 中任意元素 a、b，a*b 是 G 的元素

  例如: {-2,0,2}关于通常加法构成群，{-1, 1}关于通常乘法构成群.

- 循环群cyclic group
- 生成元g
- Z^*_p

  表示$\{a\in\{1,...,p-1\}|gcd(a,p)=1\}$，其中gcd表示最大公约数

预备知识

- 循环群cyclic group、生成元g

定义1. 设群 G 的运算记做乘法(或加法) ,如果 G 的每一个元素能写成 G 中的某个元素 a 的整数幂次(或整数倍)的形式,那么称 G 为循环群,把 a 叫做 G 的一个生成元,且把 G 记作 $\langle a\rangle$

设 $G=\langle a\rangle$,运算为乘法,单位元为 e

当 G 为无限群时, $\forall n\in\mathbb{N}^+$ 都有 $a^n\neq e$,此时

$$\langle a\rangle=\{\ldots,a^{-n},\ldots,a^{-1},e,a^1,\ldots,a^n,\ldots\}$$

称 G 为无限循环群.上述例子中 $(\mathbb{Z},+)$ 为无限循环群

当 G 为有限群时, $\exists n\in\mathbb{N}^+$ 都有 $a^n=e$,此时

$$\boxed{\langle a\rangle=\{e,a^1,\ldots,a^{n-1}\}}$$

其 $\langle a\rangle$ 阶为 n ,上述例子中 $(\mathbb{Z}_m,+)$ 是 m 阶循环群, $\mathbb{Z}_{10}^*$ 是 4 阶循环群.

<a>:循环群

a：生成园

n-1：循环群的阶

- Z^*_p

表示{a∈{1,...,p-1}|gcd(a,p)=1}，其中gcd表示最大公约数

预备知识

- Diffie-Hellman方案

方案的目的是通过一个不可信的信道，实现密钥的协商

设 $\mathcal{G}$ 是一个 PPT 算法. 输入 1^n, 输出一个循环群 $\mathbb{G}$, 阶数 为素数 q (其中 $|q| = n$), 一个 $\mathbb{G}$ 的一个生成元 g. 在 $\mathbb{G}$ 中群的运算是可以在 n 的多项式时间完成.

Algorithm 2 Diffie-Hellman 密钥交换

Input: 安全参数 1^n.

1. Alice 运行 $\mathcal{G}(1^n)$ 得到 $(\mathbb{G}, q, g)$.
2. Alice 选择均匀随机选择 $x \in \mathbb{Z}_q$, 计算 $h_1 := g^x$.
3. Alice 发送 $(\mathbb{G}, q, g, h_1)$ 给 Bob.
4. Bob 收到 $(\mathbb{G}, q, g, h_1)$. 均匀随机选择 $y \in \mathbb{Z}_q$, 计算 $h_2 := g^y$. Bob 发送 h_2 给 Alice. Bob 输出密钥 $k_B := h_1^y$.
5. Alice 收到 h_2, Alice 输出密钥 $k_A := h_2^x$.

Output: Alice 和 Bob 共享同样的密钥 g^{xy}.

Alice Bob

(1)生成(G,q,g)
(2)选择x，计算$h_1=g^x$

(3)发送(G,q,g,h_1)

(4)选择y，计算$h_2=g^y$
(5)计算$k=h_1^y=g^{xy}$

(6)发送(G,q,g,h_2)

(7)计算$k=h_2^x=g^{xy}$

公开参数：G, q, g, h_1, h_2

保密参数：x, y, g^{xy}

预备知识

- Diffie-Hellman假设

定义 (DDH 假设)

- DDH 问题: 判定性 Diffie-Hellman (Decisional Diffie-Hellman (DDH))问题是对于q 阶循环群 $\mathcal{G}$ 来说, 对于一个随机的生成元 g, 以及随机的 $x, y \leftarrow \mathbb{Z}_q$.

  $$\text{给定 } g^x, g^y \text{ 以及 } h = g^z \in \mathbb{G}\text{: 判定是否有 } h = g^{xy}.$$

- DDH 假设: 说DDH问题对于 $\mathcal{G}$ 是难解的, 如果对于随机选取的 $x, y, z \in \mathbb{Z}_q$, 任意 PPT 敌手 $\mathcal{A}$, 都存在一个可忽略函数 $\mathtt{negl}$ 使得:

$$\boxed{|\Pr[\mathcal{A}(\mathbb{G}, q, g, g^x, g^y, g^z) = 1] - \Pr[\mathcal{A}(\mathbb{G}, q, g, g^x, g^y, g^{xy}) = 1]| \leq \mathtt{negl}(\lambda)}$$

Rand　　　　Real

假设：(G,q,g,h_1,h_2)对敌手是公开的，（x,y,g^{xy}）对敌手是保密的，现在敌手想要千方百计地猜测协商密钥g^{xy}。

解释：若敌手用尽各种攻击手段，计算协商密钥g^{xy}的概率几乎等于随机乱猜，则DDH方案是安全的

定理

如果 DDH 问题对于 $\mathcal{G}$ 是难解的, 则 Diffie-Hellman 密钥交换协议对于窃听敌手是安全的.

DDH方案是有条件安全的，即存在如下假设“G上的离散对数问题是难解的”，安全性才成立。

对于方案即“根据g,x,y计算h1,h2是容易的，根据g,h1,h2计算x,y是困难的”。

预备知识

- Diffie-Hellman在OXT方案中的体现（类比）

z是保密的（相当于DH协议的y）
而y是公开的

EDBSetup(DB)

- Select key K_S for PRF F, keys K_X, K_I, K_Z for PRF F_p (with range in Z_p^*), and parse DB as $(\mathsf{ind}_i, \mathsf{W}_i)_{i=1}^d$.
- Initialize $\mathbf{T}$ to an empty array indexed by keywords from W.
- Initialize XSet to an empty set.
- For each $w \in \mathsf{W}$, build the tuple list $\mathbf{T}[w]$ and XSet elements as follows:
  - Initialize $\mathbf{t}$ to be an empty list, and set $K_e \leftarrow F(K_S, w)$.
  - For all ind in DB(w) in random order, initialize a counter $c \leftarrow 0$, then:
    * Set $\mathsf{xind} \leftarrow F_p(K_I, \mathsf{ind})$, $z \leftarrow F_p(K_Z, w \| c)$ and $y \leftarrow \mathsf{xind} \cdot z^{-1}$.
    * Compute $\mathsf{e} \leftarrow \mathsf{Enc}(K_e, \mathsf{ind})$, and append (e, y) to $\mathbf{t}$.
    * Set $\mathsf{xtag} \leftarrow g^{F_p(K_X, w)\cdot \mathsf{xind}}$ and add xtag to XSet.
  - $\mathbf{T}[w] \leftarrow \mathbf{t}$.
- $(\mathsf{TSet}, K_T) \leftarrow \mathsf{TSetSetup}(\mathbf{T})$.
- Output the key $(K_S, K_X, K_I, K_Z, K_T)$ and EDB = (TSet, XSet).

Diffie-Hellman协议

Alice　　　Bob

(1)生成(G,q,g)
(2)选择x，计算$h_1=g^x$
(3)发送(G,q,g,h_1)
(4)选择y，计算$h_2=g^y$
(5)计算$k=h_1^y=g^{xy}$
(6)发送(G,q,g,h_2)
(7)计算$k=h_2^x=g^{xy}$

$F_p(k_X,w)\cdot$xind是保密的（相当于DH协议的x），而xtag是公开的（相当于DH协议的g^{xy}）
注意：$Fp(k_X,w)$可在检索时任意构造，而xind一旦上传到Server端就不再保存

预备知识

- Diffie-Hellman在OXT方案中的体现（类比）

检索时构造满足xtoken^y=xtag的xtoken（构造h_1）

Search protocol

- The client's input is the key $(K_S, K_X, K_I, K_Z, K_T)$ and query $\bar{w} = (w_1, \ldots, w_n)$.
  It sends to the server the message (stag, xtoken[1], xtoken[2], . . .) defined as:
  - stag ← TSetGetTag(K_T, w_1).
  - For $c = 1, 2 \ldots$ and until server sends stop
    - For $i = 2, \ldots, n$, set $\boxed{\mathsf{xtoken}[c,i] \leftarrow g^{F_p(K_Z, w_1 \| c) \cdot F_p(K_X, w_i)}}$
    - Set xtoken[c] = xtoken[c, 2], . . . , xtoken[c, n].

即$g^{z1 \cdot Fp \cdot (Kx,wi)}$

- The server has input (TSet, XSet). It responds as follows.
  - It sets **t** ← TSetRetrieve(TSet, stag).
  - For $c = 1, \ldots, |\mathbf{t}|$
    - retrieve (e, y) from the c-th tuple in **t**
    - if $\forall i = 2, \ldots, n$: $\boxed{\mathsf{xtoken}[c,i]^y \in \mathsf{XSet}}$ then send e to the client.
  - When last tuple in **t** is reached, send stop to $\mathcal{C}$ and halt.
- Client sets $K_e \leftarrow F(K_S, w_1)$; for each e received, computes ind ← Dec(K_e, e) and outputs ind.

利用xtoken^{y}计算出xtag

Diffie-Hellman协议

Alice　　　　Bob

(1)生成(G,q,g)

(2)选择x，$\boxed{\text{计算}h_1=g^x}$

(3)发送(G,q,g,h_1)

(4)选择y，计算$h_2=g^y$

$\boxed{\text{(5)计算}k=h_1^{\,y}=g^{xy}}$

(6)发送(G,q,g,h_2)

(7)计算$k=h_2^{\,x}=g^{xy}$

安全性分析-关于泄露函数定义

$\mathcal{L}_{\rm oxt}(\mathsf{DB}, \mathbf{q})$ - 泄露函数
--OXT方案泄露的信息

输入：$\mathsf{DB} = (\mathsf{ind}_i, \mathsf{W}_i)_{i=1}^d$ and $\mathbf{q} = (\mathbf{s}, \mathbf{x})$
输出：$(N, \bar{\mathbf{s}}, \mathsf{SP}, \mathsf{RP}, \mathsf{IP})$

q- 查询
--包含2个关键词的AND查询

$\mathbf{q} = (\mathbf{s}, \mathbf{x})$ s:关键词w_1，x:关键词w_2

Q-查询序列
--包含多个关键词q的查询序列

N-关键词出现次数
--在所有文档中出现的关键词总数

$\bar{\mathbf{s}}$（equality pattern）-相等模式
--每个关键词出现顺序的集合，作用是判断2个q是否有相同的s-term

We represent a sequence of Q non-adaptive 2-conjunction queries by $\mathbf{q} = (\mathbf{s}, \mathbf{x})$ where an individual query is a 2-term conjunction $\mathbf{s}[i] \wedge \mathbf{x}[i]$ which we write as $\mathbf{q}[i] = (\mathbf{s}[i], \mathbf{x}[i])$. $\mathcal{L}_{\rm oxt}(\mathsf{DB}, \mathbf{q})$ gets $\mathsf{DB} = (\mathsf{ind}_i, \mathsf{W}_i)_{i=1}^d$ and $\mathbf{q} = (\mathbf{s}, \mathbf{x})$ as input and outputs $(N, \bar{\mathbf{s}}, \mathsf{SP}, \mathsf{RP}, \mathsf{IP})$, which are defined below.

- $N = \sum_{i=1}^d |\mathsf{W}_i|$ is the total number of appearances of keywords in documents.
- $\bar{\mathbf{s}} \in [m]^Q$ is the *equality pattern* of $\mathbf{s} \in \mathsf{W}^Q$ indicating which queries have the equal s-terms. Formally, $\bar{\mathbf{s}} \in [m]^Q$ is formed by assigning each keyword an integer in $[m]$ determined by the order of appearance in $\mathbf{s}$. For example, if $\mathbf{s} = (a, a, b, c, a, c)$ then $\boxed{\bar{\mathbf{s}} = (1, 1, 2, 3, 1, 3)}$. To compute $\bar{\mathbf{s}}[i]$ one finds the least j such that $\mathbf{s}[j] = \mathbf{s}[i]$ and then lets $\bar{\mathbf{s}}[i] = |\{\mathbf{s}[1], \ldots, \mathbf{s}[j]\}|$ be the number of unique keywords appearing at indices less than or equal to j.

安全性分析-关于泄露函数定义

SP（Size Pattern）-数量模式
--检索关键词w1返回的文档数量

RP（results pattern）-结果模式
--多关键词检索返回的文档索引集合

IP（conditional intersetion pattern）
-条件交集模式
--IP是维护关于关键词两两查询交集的一张表

- SP is the *size pattern* of the queries, which is the number of documents matching the first keyword in each query. Formally, $\mathsf{SP} \in [d]^Q$ and $\mathsf{SP}[i] = |\mathsf{DB}(\mathbf{s}[i])|$.
- RP is the *results pattern* of the queries, which are the indices of documents matching the entire conjunction. Formally, RP is vector of size Q with $\mathsf{RP}[i] = \mathsf{DB}(\mathbf{s}[i]) \cap \mathsf{DB}(\mathbf{x}[i])$ for each i.
- IP is the *conditional intersection pattern*, which is formally a Q by Q table defined by

$$\mathsf{IP}[i,j] = \begin{cases} \mathsf{DB}(\mathbf{s}[i]) \cap \mathsf{DB}(\mathbf{s}[j]) & \text{if } i \neq j \text{ and } \mathbf{x}[i] = \mathbf{x}[j] \\ \emptyset & \text{otherwise} \end{cases}$$

例如q1=(w1,w2),q2=(w2,w3)， 则q1∩q2=(w1,w2,w3)
则x[i]=x[j]=w2， IP[i,j]表示q1查询及q2查询返回文档索引的交集，即返回满足(w1,w2,w3)的索引集合

安全性分析-泄露内容（方案的局限性）

- N：表示EDB允许泄露信息的上界。
- $\overline{\mathbf{s}}$：泄露了多次查询中s-term的查询次数。
- SP：泄露了单个查询中满足s-term（w1）的文档数量
- RP：是查询的结果并没有泄露任何额外的信息
- IP：对于2个s-term不同x-term相同的查询，那么如果有一个文档满足两个s-term，那么匹配两个s-term的索引集就会被泄露

安全性分析-OXT的安全性证明

- 证明思路：

最终目的是证明boolean查询在自适应方案下的安全性，首先通过非自适应性方案下关键词两两连接特例的安全性进行证明。特例证明基本足以覆盖最终证明的难点。

OXT是语义安全的，能够抵抗2个连词的非适应性攻击

Theorem 5 *Let $\mathcal{L}_{\text{oxt}}$ be as defined above, and suppose that the T-set implementation Σ from Section 6. Then SSE scheme* OXT *is $\mathcal{L}_{\text{oxt}}$-semantically-secure against non-adaptive attacks where are all queries are 2-conjunctions, assuming that the DDH assumption holds in G, that F and F_p are secure PRFs, that* (Enc, Dec) *is an IND-CPA secure symmetric encryption scheme, and the conditions from Theorem 7 hold.*

PRF Security. Let X and Y be sets, and let $F : \{0,1\}^\lambda \times X \to Y$ be a function. We say that F is a *pseudorandom function (PRF)* if for all efficient adversaries A, $\mathbf{Adv}^{\text{prf}}_{F,A}(\lambda)$ is negligible, where

$$\mathbf{Adv}^{\text{prf}}_{F,A}(\lambda) = \Pr[A^{F(K,\cdot)}(1^\lambda) = 1] - \Pr[A^{f(\cdot)}(1^k) = 1]$$

where the probability is over the randomness of A, $K \xleftarrow{\$} \{0,1\}^\lambda$, and $f \xleftarrow{\$} \mathsf{Fun}(X,Y)$.

前置条件1：伪随机函数(PRF)存在

Lemma 3 *Suppose the DDH assumption holds for in G. Then, for any integers α, β (polynomial in λ) any efficient adversary A, we have*

$$\Pr[A(g, g^{\mathbf{a}}, g^{\mathbf{b}}, g^{\mathbf{a}\mathbf{b}^{\mathrm{T}}}) = 1] - \Pr[A(g, g^{\mathbf{a}}, g^{\mathbf{b}}, \mathbf{M}) = 1] \le \text{neg}(\lambda),$$

where $\mathbf{a}$ is uniform over $(Z_p^*)^\alpha$, $\mathbf{b}$ is uniform over $(Z_p^*)^\beta$, and $\mathbf{M}$ is uniform over $G^{\alpha\times\beta}$.

前置条件2：DDH假设成立

Lemma 4 *For every adversary A there exists adversaries B and B' which run in essentially the same time as A, such that*

$$\Pr[\mathbf{Cor}^{\mathsf{OXT}}_A(\lambda) = 1] \le 2 \cdot \mathbf{Adv}^{\text{prf}}_{F_p,B}(\lambda) + \mathbf{AdvCor}^{\Pi}_{B'}(\lambda) + N^2/(p-1) + N/p,$$

where $N = \Sigma_{i=1}^{d} |\mathsf{W}_i|$ is the total number of appearances of keywords in all documents, p is the order of the group G, and Π is the T-set implementation.

前置条件3：引理4得证（证明过程见附录）