Chapter 1

Introduction

For traditional encryption schemes, security is based on indistinguishability (IND) of ciphertexts [18] or equivalently semantic security [17]. For schemes that are secure in this setting, the adversary should learn no information from seeing a ciphertext. However, this also means that one cannot process encrypted data efficiently. For instance, keyword search can only be achieved in linear time [9, 28].

Schemes allowing for better processing of encrypted data are devised but the ciphertext often possess some additional properties, thus leaking information about the underlying plaintext. Those schemes usually do not satisfy the usual indistinguishability notion of security because the additional properties can be exploited to generate attacks. In that case, one needs to understand what is the maximum level of security those schemes can offer. For our interest, we will focus on a specific encryption scheme of the type, known as deterministic encryption.

1.1 Deterministic Encryption

In CRYPTO 2007, Bellare et al. presented a construction of deterministic encryption (DE) scheme based on (randomized) public-key encryption schemes, and defined security notion for their scheme in the random oracle model [4]. In the follow-up papers, definitional equivalences of security notion without random oracles are studied [5, 8].

In this section, we will describe the key ideas in the construction, and analyse security of the scheme under the security notion defined in the original papers.

Key generation	Encryption(pk, x)	Decryption(pk, sk, y)
1 : $(\mathsf{pk}, \mathsf{sk}) \leftarrow_{\$} \mathcal{K}(1^n)$	1 : $\omega \leftarrow H(\mathsf{pk} \| x)$	1 : $x \leftarrow \mathcal{D}(\mathsf{sk}, y)$
	2 : $y \leftarrow \mathcal{E}(\mathsf{pk}, x; \omega)$	2 : $\omega \leftarrow H(\mathsf{pk} \| x)$
	3 : **return** y	3 : **if** $\mathcal{E}(\mathsf{pk}, x; \omega) = y$ **return** x
		4 : **return** $\perp$

Figure 1.1: Deterministic encryption based on hashing

1.1.1 Construction based on Hash function

Let $(\mathcal{K}, \mathcal{E}, \mathcal{D})$ be any randomized public-key encryption scheme, and $H(\cdot)$ a hash function. The idea of deterministic encryption is that instead of instead of letting the encryption algorithm $\mathcal{E}$ to determine its randomness, we will use the hash function $H(\cdot)$ to flip the coins deterministically.

For the construction to work, of course we demand $H(\mathsf{pk} \| x) \in \mathrm{Coin}_{\mathsf{pk}}(|x|)$, where $\mathrm{Coin}_{\mathsf{pk}}$ denotes the set of coins of $\mathcal{E}(\mathsf{pk}, x)$. Intuitively, this scheme is secure because for an adversary to recover the message, he effectively has to know the underlying randomness used by the encryption algorithm, but for a secure hash function the adversary cannot do so.

1.1.2 Generalization of Previous Construction

The construction above can be generalised. Instead of hashing, we can use trapdoor permutation to generate the coins. We shall first define trapdoor permutation formally.

Definition 1 (Trapdoor permutation) *A trapdoor function is a family of functions that is easy to compute but hard to invert. However, if given some additional information, known as the 'trapdoor', the inversion can be computed efficiently.*
Formally, let $F = \{f_k : D_k \to R_k\}(k \in K)$ be a collection of one-way functions, then F is a trapdoor function if the following holds:

- *There exists a probabilistic polynomial time (PPT) sampling algorithm Gen such that $Gen(1^n) = (k, t_k)$ and $t_k \in \{0,1\}^*$ satisfies that $|t_k|$ is polynomial in length. Each t_k is called the trapdoor corresponding to k.*
- *Given input k, there exist a PPT algorithm that outputs $x \in D_k$.*
- *For any $k \in K$, there exist a PPT algorithm that computes f_k correctly.*
- *For any $k \in K$, there exist a PPT algorithm A such that $y = A(k, f_k(x), t_k)$, and $f_k(y) = f_k(x)$. That is, the function can be inverted efficiently with the trapdoor.*

Key generation	Encryption(pk, x)	Decryption(pk, sk, y)
1 : $(\phi, \tau) \leftarrow_{\$} \mathcal{G}(1^n)$	1 : $(\phi, \bar{\mathsf{pk}}, p) \leftarrow \mathsf{pk}$	1 : $(\tau, \bar{\mathsf{sk}}) \leftarrow \mathsf{sk}$
2 : $s \leftarrow_{\$} \{0,1\}^n$	2 : $y \leftarrow F(\phi, x)$	2 : $y \leftarrow \mathcal{D}(\bar{\mathsf{sk}}, c)$
3 : $(\bar{\mathsf{pk}}, \bar{\mathsf{sk}}) \leftarrow_{\$} \mathcal{K}(1^n)$	3 : $\omega \leftarrow GetCoins(F, \phi, x, s)$	3 : $x \leftarrow \bar{F}(\tau, y)$
4 : $\mathsf{pk} \leftarrow (\phi, \bar{\mathsf{pk}}, s)$	4 : $c \leftarrow \mathcal{E}(\mathsf{pk}, y; \omega)$	4 : **return** x
5 : $\mathsf{sk} \leftarrow (\tau, \bar{\mathsf{sk}})$	5 : **return** c	
6 : **return** $(\mathsf{pk}, \mathsf{sk})$		

Figure 1.2: Deterministic encryption based on trapdoor permutations

- *For any $k \in K$, without the trapdoor t_k, the adversary of any PPT adversary*

$$\mathsf{Adv}^{\text{trapdoor}}_{\mathcal{A},Trapdoor}(n) = \Pr\left[\begin{array}{c}(t, t_k) \leftarrow Gen(1^n), x \leftarrow D_k, y \leftarrow f_k(x), \\ z \leftarrow \mathcal{A}(1^n, k, y) : f_k(x) = f_k(y)\end{array}\right]$$

  is negligible.

For ease of notation, we write trapdoor permutation as $(G, F, \bar{F})$, where G is the key generation algorithm, F is the trapdoor permutation, and $\bar{F}$ is the inverse. Further, we write $F^n(\cdot)$ to denote $F(F(\cdots F(\cdot)))$, that is F is applied to the input n times. Finally, we denote $GetCoins(F, \phi, \cdot, \cdot)$ to be a pseudo-random generator based on one-way function F and its public key ϕ. Such construction can be found in the [7, 32, 16].

The generalised deterministic encryption scheme has a very similar construction to the case where hash function is used. The main differences are: (1) instead of encrypting the plaintext straight away, the trapdoor permutation is applied to the plaintext before using the probabilistic encryption algorithm $\mathcal{E}$, and (2) the trapdoor function is used to generate the randomness for the encryption scheme.

1.1.3 Usefulness of Deterministic Encryption

There are numerous searchable encryption schemes with strong security guarantees include [9, 19, 1, 3, 11, 10] in the public-key setting, and [28, 15, 13] in the symmetric-key setting. However, all the schemes require linear search time, which is not ideal for large databases.

On the other hand, researchers have proposed sub-linear time searchable encryption in [27, 2, 20, 14, 22, 23, 25, 21, 12, 30]. However, security of these schemes are usually not analysed, which means that they can be vulnerable to various type of attacks. Deterministic encryption is one of the first schemes that is efficient in encrypting database and making queries, with provable security guarantees. In particular, for deterministic encryption, searching for equality can be done in log-time with binary search [31], or log-log-time with more advanced Van Emde Boas tree [29]. In our research, we will focus mainly on the database application of deterministic encryption.

Experiment $\mathrm{EXP}_{\mathrm{I}}^{\mathrm{IND}}(n)$

$st \leftarrow I_c(1^n)$

$(m_0, m_1) \leftarrow I_m(st)$

$b \leftarrow_{\$} \{0, 1\}$

$c \leftarrow \mathrm{Encryption}(pk, m_b)$

$b' \leftarrow I_g(\mathsf{pk}, c, st)$

Figure 1.3: IND game for deterministic encryption

1.1.4 Security of Deterministic Encryption

SECURITY NOTION IN THE ORIGINAL PAPER. As the scheme is deterministic, it is impossible to meet classic IND-CPA security [18]. In the original papers for deterministic encryption [4, 5, 5], the authors presented a set of security notions based on the traditional semantic security and IND security with a less powerful adversary, and proved equivalence of the notions. For our interest, we will present the IND adversary.

An IND adversary is a triple $I = (I_c, I_m, I_g)$ of PPT algorithms. I_c is an algorithm that, given the security parameter as the input, generates a state that will be used by I_m. I_m then outputs two plaintexts (m_0, m_1) given the state passed on by I_c. The challenger picks a bit b randomly from $0, 1$ and encrypts m_b. The encrypted message is then sent back to the adversary. Finally, the adversary guesses the bit b by running I_g with public key, the encryption of m_b, and the state generated by I_c.

We say that a scheme is IND secure, if for any IND-adversary I, the probability that I guesses the bit b right is only negligibly higher than half. The adversary presented here has a few key differences to the traditional IND-adversary:

- The algorithms I_c and I_m accessed by the adversary have no access to the public key.
- The final algorithm I_g only has access to the state generated by I_c, instead of the plaintexts (m_0, m_1).
- The message space has high min-entropy [24].

All these assumptions on the power of the adversary are very strong. In practice, one can hardly expect the adversary to gain access to the public key only at the guessing stage. It is also unlikely for an adversary to not have the knowledge of the plaintexts when he makes the guess. Lastly, in many applications, the message space in fact has very low entropy. Hence, security in this notion does not necessarily imply security in practice.

SECURITY NOTION OF IND-DCPA. Alternatively, IND-DCPA (indistinguishability under distinct chosen-plaintext attack) is defined in [6]. As deterministic encryption will always

leak plaintext equality, the idea of IND-DCPA is that the adversary should only make distinct queries to the oracle.

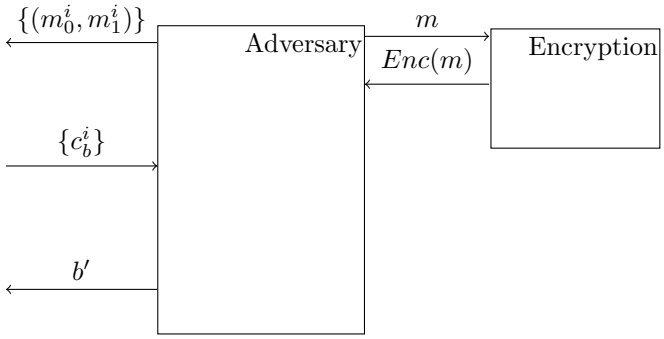

Figure 1.4: Cryptographic game of IND-DCPA

In the game, the adversary queries the challenger pairs of messages $(m_0^1, m_1^1), (m_0^2, m_1^2), \cdots, (m_0^q, m_1^q)$, where $m_b^1, \cdots, m_b^q$ are all distinct for $b \in \{0, 1\}$. The challenger then randomly picks $b \in \{0, 1\}$, encrypts all m_b^i for $i \in \{1, \cdots, q\}$ and sends them back to the adversary. The adversary, with the help of the encryption oracle, has to guess the bit b. He returns $b^{'}$ as his guess and he wins if $b = b^{'}$.

Just like the security notion in the original paper, assumption of IND-DCPA is hard to meet in practice. In particular, for databases, there is no reason why the same plaintext cannot be encrypted twice.

1.2 Attack on Database Encrypted using Deterministic Encryption

Deterministic encryption has fairly strong security guarantees if the assumptions of the security notion are met. However, for database applications, the assumptions are impossible to achieve. In [26], Naveed et al. have proposed frequency attack on databases encrypted using deterministic encryption, with auxiliary information on the distribution of the plaintext. They have tested their attack on National Inpatient Sample (NIS) database of the Healthcare Cost and Utilization Project (HCUP). In the experiment, they are able to recover almost all information of the database.

1.2.1 Notation

We denote $c = (c_0, c_1, \cdots, c_n)$ to be the list of entries for a column in the database. We write $Hist(c)$ to be the function that generates the histogram of c. On input c_i, $Hist(c)$ will output the frequency of c_i inside c. Further, we write $vSort(\cdot)$ to be the function that sorts a histogram in decreasing order of frequency. So $vSort(Hist(c))[0]$ will be the element in c that has the highest frequency. Finally, we define $Rank_\psi(c_i)$ to be the rank of c_i in the sorted histogram ψ. That is, if c_i is the most frequent ciphertext in c, $Rank_{vSort(Hist(c))}(c_i) = 0$.

1.2.2 The attack

The attack used in the paper against deterministically encrypted databases is the most basic and famous attack, known as frequency attack. The attack relies on auxiliary information on the plaintext. Let c be the target of attack, and z be some auxiliary dataset. The attack can be written as follows:

Attack(z, c)

1 : compute $\psi \leftarrow vSort(Hist(c))$
2 : compute $\pi \leftarrow vSort(Hist(z))$
3 : $output A : C_k \rightarrow M_k$ such that
$$A(c) = \pi[Rank_\psi(c)]$$

The attack does no more than just assigning the most frequent plaintext of the auxiliary dataset to the most frequent ciphertext in the target database. But for statistical databases, the distributions usually do not change, so the attack has a good chance of recovering the plaintexts.

On a side note, additional datasets may be very easy to obtain in some cases. In the paper that the attack described is based on, the auxiliary data is the Texas Inpatient Public Use Data File (PUDF), which is publicly available online. Usage of the database only requires the user to sign a data use agreement, but there is no reason to assume that an evil adversary will keep his promise.

1.3 Discussion

In this chapter, we have shown that deterministic encryption is very useful in encrypting databases. However, for such application, the assumptions of the security notion introduced in [4, 8, 5] cannot be met, so attacks such as frequency attack [26] can be deployed to exploit the cryptosystem.

As the security notions defined in the original paper are not useful against statistical attacks, we wish to construct new security notions that have security guarantee against such attacks. Furthermore, we want to find schemes that are efficient in encryption and database queries, and prove their security in the new security notion.

Bibliography

[1] Michel Abdalla, Mihir Bellare, Dario Catalano, Eike Kiltz, Tadayoshi Kohno, Tanja Lange, John Malone-Lee, Gregory Neven, Pascal Paillier, and Haixia Shi. *Searchable Encryption Revisited: Consistency Properties, Relation to Anonymous IBE, and Extensions*, pages 205–222. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[2] Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, and Yirong Xu. Order preserving encryption for numeric data. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD '04, pages 563–574, New York, NY, USA, 2004. ACM.

[3] Joonsang Baek, Reihaneh Safavi-Naini, and Willy Susilo. *Public Key Encryption with Keyword Search Revisited*, pages 1249–1259. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[4] Mihir Bellare, Alexandra Boldyreva, and Adam O'Neill. *Deterministic and Efficiently Searchable Encryption*, pages 535–552. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[5] Mihir Bellare, Marc Fischlin, Adam O'Neill, and Thomas Ristenpart. *Deterministic Encryption: Definitional Equivalences and Constructions without Random Oracles*, pages 360–378. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[6] Mihir Bellare, Tadayoshi Kohno, and Chanathip Namprempre. Authenticated encryption in ssh: Provably fixing the ssh binary packet protocol. In *Proceedings of the 9th ACM Conference on Computer and Communications Security*, CCS '02, pages 1–11, New York, NY, USA, 2002. ACM.

[7] Manuel Blum and Silvio Micali. How to generate cryptographically strong sequences of pseudorandom bits. *SIAM Journal on Computing*, 13(4):850–864, 1984.

[8] Alexandra Boldyreva, Serge Fehr, and Adam O'Neill. *On Notions of Security for Deterministic Encryption, and Efficient Constructions without Random Oracles*, pages 335–359. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[9] Dan Boneh, Giovanni Di Crescenzo, Rafail Ostrovsky, and Giuseppe Persiano. *Public Key Encryption with Keyword Search*, pages 506–522. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[10] Dan Boneh and Brent Waters. *Conjunctive, Subset, and Range Queries on Encrypted Data*, pages 535–554. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[11] Xavier Boyen and Brent Waters. *Anonymous Hierarchical Identity-Based Encryption (Without Random Oracles)*, pages 290–307. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[12] R. Brinkman, L. Feng, J. Doumen, P. H. Hartel, and W. Jonker. Efficient tree search in encrypted data. *Information Systems Security*, 13(3):14–21, 2004.

[13] Yan-Cheng Chang and Michael Mitzenmacher. *Privacy Preserving Keyword Searches on Remote Encrypted Data*, pages 442–455. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[14] Ernesto Damiani, S. De Capitani Vimercati, Sushil Jajodia, Stefano Paraboschi, and Pierangela Samarati. Balancing confidentiality and efficiency in untrusted relational dbmss. In *Proceedings of the 10th ACM Conference on Computer and Communications Security*, CCS '03, pages 93–102, New York, NY, USA, 2003. ACM.

[15] Eu-Jin Goh. Secure indexes. Cryptology ePrint Archive, Report 2003/216, 2003. `http://eprint.iacr.org/2003/216/`.

[16] O. Goldreich and L. A. Levin. A hard-core predicate for all one-way functions. In *Proceedings of the Twenty-first Annual ACM Symposium on Theory of Computing*, STOC '89, pages 25–32, New York, NY, USA, 1989. ACM.

[17] Shafi Goldwasser and Silvio Micali. Probabilistic encryption & how to play mental poker keeping secret all partial information. In *Proceedings of the Fourteenth Annual ACM Symposium on Theory of Computing*, STOC '82, pages 365–377, New York, NY, USA, 1982. ACM.

[18] Shafi Goldwasser and Silvio Micali. Probabilistic encryption. *Journal of Computer and System Sciences*, 28(2):270 – 299, 1984.

[19] Philippe Golle, Jessica Staddon, and Brent Waters. *Secure Conjunctive Keyword Search over Encrypted Data*, pages 31–45. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[20] Hakan Hacigümüş, Bala Iyer, Chen Li, and Sharad Mehrotra. Executing sql over encrypted data in the database-service-provider model. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, SIGMOD '02, pages 216–227, New York, NY, USA, 2002. ACM.

[21] Hakan Hacıgümüş, Bala Iyer, and Sharad Mehrotra. *Efficient Execution of Aggregation Queries over Encrypted Relational Databases*, pages 125–136. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[22] Bijit Hore, Sharad Mehrotra, and Gene Tsudik. A privacy-preserving index for range queries. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pages 720–731. VLDB Endowment, 2004.

[23] Bala Iyer, Sharad Mehrotra, Einar Mykletun, Gene Tsudik, and Yonghua Wu. *A Framework for Efficient Storage Security in RDBMS*, pages 147–164. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[24] R. Koenig, R. Renner, and C. Schaffner. The operational meaning of min- and max-entropy. *ArXiv e-prints*, July 2008.

[25] Jun Li and Edward R. Omiecinski. *Efficiency and Security Trade-Off in Supporting Range Queries on Encrypted Databases*, pages 69–83. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[26] Muhammad Naveed, Seny Kamara, and Charles V. Wright. Inference attacks on property-preserving encrypted databases. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 644–655, New York, NY, USA, 2015. ACM.

[27] Gultekin Ozsoyoglu, David A. Singer, and Sun S. Chung. *Anti-Tamper Databases*, pages 133–146. Springer US, Boston, MA, 2004.

[28] Dawn Xiaodong Song, David Wagner, and Adrian Perrig. Practical techniques for searches on encrypted data. In *Proceedings of the 2000 IEEE Symposium on Security and Privacy*, SP '00, pages 44–, Washington, DC, USA, 2000. IEEE Computer Society.

[29] P. van Emde Boas. Preserving order in a forest in less than logarithmic time and linear space. *Information Processing Letters*, 6(3):80 – 82, 1977.

[30] Hui Wang and Laks V. S. Lakshmanan. Efficient secure query evaluation over encrypted xml databases. In *Proceedings of the 32Nd International Conference on Very Large Data Bases*, VLDB '06, pages 127–138. VLDB Endowment, 2006.

[31] Louis F. Williams, Jr. A modification to the half-interval search (binary search) method. In *Proceedings of the 14th Annual Southeast Regional Conference*, ACM-SE 14, pages 95–101, New York, NY, USA, 1976. ACM.

[32] A. C. Yao. Theory and application of trapdoor functions. In *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*, pages 80–91, Nov 1982.