Security Analysis for an Order Preserving Encryption Scheme

Liangliang Xiao I-Ling Yen Dongdai Lin

Abstract

The development of third-party hosting, IT outsourcing, service clouds, etc. raises important security concerns. It would be safer to encrypt critical data hosted by a third-parity, but meanwhile, the database should be able to process queries on encrypted data. Many research works have been developed to support search query processing on encrypted data, including the order preserving encryption (OPE) schemes.

Security analysis plays an important role on secure algorithm design. It can help understand the level of security assurance of the algorithm. Currently, security analysis for OPE schemes is limited. In [8], a cryptographic-based OPE scheme, $SE_{m,n}$, has been proposed. It defines the ideal model and the real model and construct the OPE in the real model to satisfy the idea model security. However, the security of the idea model itself has not been analyzed. In this paper, we first analyze the information leaks in $SE_{m,n}$, and then use information theory to analyze the security of $SE_{m,n}$. More specifically, we derive an upper bound on the probability for an adversary to recover the plaintext encrypted by $SE_{m,n}$.

Key Words: Order preserving encryption, hypergeometric distribution, information entropy, chosen plain text attacks.

1 Introduction

It has been a common practice for companies to outsource their online business logics to Web hosting service providers for over a decade. Generally, this involves the storage of databases which potentially contain sensitive information as well as the execution of access logics to the databases. The cloud computing further pushes forward this paradigm and creates a whole spectrum of third-party hosting business, from "database as a service (DAS)" [3, 2, 10] to full-scale service hosting. With the many benefits of outsourcing, such as reduced computation and personnel management costs, the security concerns emerge. For example, if the hosting service provider is compromised, the attacker can retrieve the sensitive data of the client companies. Or if there is a change in management of the hosting company, such as reorganization or buyout [14], the potential threat

increases due to the additional exposure to multiple management personnel and the unestablished policies regarding the handling of critical information in such situations.

The security problems with the outsourced databases can be solved if the sensitive data are encrypted. Naturally it leads to the problem that how the database management system (DBMS) can process queries on encrypted data. Classical cryptographic encryption algorithms encrypt a plain text to a cipher text such that no probabilistic polynomial time (PPT) algorithm can derive any information of the plain text from the cipher text. This property makes it infeasible for the DBMS to perform search on the cipher texts and, thus, infeasible to process most of the queries. To cope with the problem, various encryption algorithms and schemes have been proposed to support different types of searches on encrypted data [1, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 15, 16, 17, 18].

Search queries can be classified into extact-match search queires and range search queries. Range queires are more general and more difficult to handle under security constraints. One important class of methods to enable range query processing on encrypted data is order preserving encryption (OPE) [4, 11, 12, 1, 8]. OPE is a symmetric-key encryption scheme that preserves the order of the data. Thus, range queries can be handled efficiently using conventional DBMS techniques, such as establishing the B+ tree on cipher texts. There are various constructions of OPE. In [4], the proposed OPE first generates a sequence of random numbers and then, encrypts an integer x by adding x to the sum of the first x random numbers. In [11], a sequence of strictly increasing polynomial functions are used to construct the OPE. The encryption of an integer x is the outcome of the iterative operations of those functions on x. In [12], the OPE is constructed by using a mapping function composed of partition and identification functions. The partition function divides the range into multiple partitions, and the identification function assigns an identifier to each partition. Then, the mapping function maps an integer x to an identifier. Since different integers may be mapped to the same identifier, the OPE may output false comparison results. In [1], the authors construct the OPE in three steps: modeling the input and target distributions, flattening the plaintext database into a flat database, and transforming the flat database into the cipher database.

Security analysis for the OPE algorithms has not been widely investigated. In [1], the authors construct an OPE and analyze its security, but the analysis has some limitations: (1) It assumes that the adversaries can only view cipher texts. (2) The analysis is based on experiments, i.e. the output of OPE and the target distribution cannot be distinguished by Kolmogorov-Smirnov test, not based on cryptographic analysis. [8] initiates the cryptographic study of OPE. It defines the security of an OPE algorithm using the ideal model and the real model, and construct the OPE in the real model to satisfy the security "implied" in the idea model. Although the OPE construction meets the "best-possible" security notions of the idea model, the security of the idea model itself has not been analyzed [8].

In this paper we analyze the security of the OPE scheme, $SE_{m,n}$, constructed in [8]. We first show that there are information leaks in $SE_{m,n}$ under chosen plain text attacks. Then, we estimate the degree of information leaks and derive an upper bound on the probability for the adversary to recover the plaintext encrypted using $SE_{m,n}$. To estimate the probability, we analyze the lower bound of the average information entropy of the hypergeometric distributions which have

the intimate relation with $SE_{m,n}$. Based on the lower bound estimation, we compute the upper bound on the plaintext compromising probability of $SE_{m,n}$ for the special situation that there is no chosen plain text attacks. Finally, we extend the special case and compute the upper bound on the plaintext compromising probability for the case when the adversary does have knowledge of some plaintext and ciphertext pairs.

The rest of the paper is organized as follows. In Section 2, the lower bound of the average information entropy of the hypergeometric distributions is computed. In Section 3, we discuss the problem on information leaks in $SE_{m,n}$ under chosen plain text attacks. In Section 4, we derive the upper bound on the probability for the adversary to recover the plain texts encrypted by $SE_{m,n}$ under chosen plain text attacks. Finally, we conclude the paper in Section 5.

2 Lower Bound of the Average Information Entropy of Hypergeometric Distribution

In the OPE scheme $SE_{m,n}$ introduced in [8], a plain text x is mapped to its cipher by a "binary-search-like" process in the ciphper space with the searched points being mapped back to the plain text space using the hypergeometric distribution. Our goal is to derive the upper bound on the probability for the adversary to recover the plain text x in $SE_{m,n}$. To achieve the goal, we need to analyze the information entropy of the plain text x that can remain secure under chosen plain text attacks. We first consider the special case when there is no chosen plain text attacks. The information entropy of a plain text for this case is essentially the information entropy of the hypergeomeric distribution. Thus, in this section, we derive the lower bound of the average information entropy of hypergeometric distributions. Before showing the lower bound, we first prove two lemmas.

Lemma 2.1 *$\forall\ \epsilon > 0$, $\exists\ x_\epsilon > 0$, such that $(1+\frac{1}{x})^x = e + \frac{c_1}{x}$ where $c_1 \in [-\frac{e}{2} - \epsilon, -\frac{e}{2} + \epsilon]$ for $|x| \geq x_\epsilon$.*

Proof. $(1+\frac{1}{x})^x = e^{x\cdot\ln(1+\frac{1}{x})} = e^{x\cdot(\frac{1}{x}-\frac{1}{2x^2}+o(\frac{1}{x^2}))} = e^{1-\frac{1}{2x}+o(\frac{1}{x})} = e\cdot e^{-\frac{1}{2x}+o(\frac{1}{x})} = e\cdot(1-\frac{1}{2x}+o(\frac{1}{x}) + o(-\frac{1}{2x}+o(\frac{1}{x}))) = e - \frac{e}{2x} + o(\frac{1}{x}) = e + (-\frac{e}{2}+o(1))\cdot\frac{1}{x}$. $\square$

Lemma 2.2 *$\forall\ \epsilon > 0$, $\exists\ 0 < c_\epsilon < 1$ and $y_\epsilon > 0$, such that $\frac{c_\epsilon\cdot(\frac{ex}{y})^y}{\sqrt{2\pi y}} \leq \binom{x}{y} \leq \frac{e^\epsilon\cdot(\frac{ex}{y})^y}{\sqrt{2\pi y}}$ for $x \geq y^2$ and $y \geq y_\epsilon$.*

Proof. *According to Stirling's formula, $x! = \sqrt{2\pi x}(\frac{x}{e})^x e^{\lambda_x}$ where $\frac{1}{12x+1} < \lambda_x < \frac{1}{12x}$. Hence $\binom{x}{y} = \frac{x!}{y!\cdot(x-y)!} = \frac{\sqrt{2\pi x}(\frac{x}{e})^x e^{\lambda_x}}{\sqrt{2\pi y}(\frac{y}{e})^y e^{\lambda_y}\cdot\sqrt{2\pi(x-y)}(\frac{x-y}{e})^{x-y}e^{\lambda_{x-y}}} = \frac{\sqrt{\frac{x}{x-y}}\cdot(\frac{x}{y})^y\cdot(\frac{x}{x-y})^{x-y}}{\sqrt{2\pi y}}\cdot e^{\lambda_x-\lambda_y-\lambda_{x-y}} = \frac{(\frac{x}{y})^y\cdot(\frac{x}{x-y})^{x-y+\frac{1}{2}}}{\sqrt{2\pi y}}\cdot e^{\lambda_x-\lambda_y-\lambda_{x-y}}$.*

According to Lemma 2.1, $(\frac{x}{x-y})^{x-y+\frac{1}{2}} = (1+\frac{y}{x-y})^{x-y+\frac{1}{2}} = (e+\frac{c_1 y}{x-y})^{\frac{y}{x-y}\cdot(x-y+\frac{1}{2})} = (e + \frac{c_1 y}{x-y})^{y\cdot(1+\frac{1}{2(x-y)})} = e^{y\cdot(1+\frac{1}{2(x-y)})}\cdot(1+\frac{c_1 y}{e(x-y)})^{y\cdot(1+\frac{1}{2(x-y)})} = e^{y\cdot(1+\frac{1}{2(x-y)})}\cdot(e+\frac{c_1^2 y}{e(x-y)})^{\frac{c_1 y}{e(x-y)}\cdot y\cdot(1+\frac{1}{2(x-y)})} =$

$e^{y\cdot(1+\frac{1}{2(x-y)})}\cdot(e+\frac{c_1^2 y}{e(x-y)})^{\frac{c_1 y^2\cdot(1+\frac{1}{2(x-y)})}{e(x-y)}}$ *for* $|\frac{x-y}{y}|\geq x_\epsilon$ *and* $|\frac{e(x-y)}{c_1 y}|\geq x_\epsilon$.

For $\epsilon>0$, $\exists\ x'_\epsilon>0$ *such that* $e^y<e^{y\cdot(1+\frac{1}{2(x-y)})}=e^{y+\frac{y}{2(x-y)}}\leq e^{y+\epsilon}$ *and* $e\leq e+\frac{c_1^2 y}{e(x-y)}=e+\frac{c_1^2\cdot\frac{y}{x}}{e(1-\frac{y}{x})}\leq e+\epsilon$ *for* $\frac{x}{y}\geq x'_\epsilon$ *and* $\frac{(-\frac{e}{2}-\epsilon)\cdot(1+\epsilon)}{e(1-\epsilon)}\leq\frac{c_1 y^2(1+\frac{1}{2(x-y)})}{e(x-y)}=\frac{c_1\frac{y^2}{x}(1+\frac{1}{2(x-y)})}{e(1-\frac{y}{x})}\leq 0$ *for* $x-y\geq x'_\epsilon$, $\frac{x}{y}\geq x'_\epsilon$, *and* $x\geq y^2$. *Therefore* $e^y\cdot(e+\epsilon)^{\frac{(-\frac{e}{2}-\epsilon)\cdot(1+\epsilon)}{e(1-\epsilon)}}\leq(\frac{x}{x-y})^{x-y+\frac{1}{2}}\leq e^{y+\epsilon}$ *for* $|\frac{x-y}{y}|\geq x_\epsilon$, $|\frac{e(x-y)}{c_1 y}|\geq x_\epsilon$, $x-y\geq x'_\epsilon$, $\frac{x}{y}\geq x'_\epsilon$, *and* $x\geq y^2$. *Also* $\exists\ x''_\epsilon>0$ *such that* $1-\epsilon\leq e^{\lambda_x-\lambda_y-\lambda_{x-y}}<1$ *for* $x\geq x''_\epsilon, y\geq x''_\epsilon$, *and* $x-y\geq x''_\epsilon$.

For $x_\epsilon, x'_\epsilon, x''_\epsilon$, $\exists\ y_\epsilon>0$ *such that* $|\frac{x-y}{y}|\geq x_\epsilon$, $|\frac{e(x-y)}{c_1 y}|\geq x_\epsilon$, $x-y\geq x'_\epsilon$, $\frac{x}{y}\geq x'_\epsilon$, $x\geq x''_\epsilon, y\geq x''_\epsilon$, $x-y\geq x''_\epsilon$ *for* $x\geq y^2$ *and* $y\geq y_\epsilon$. *Hence* $\forall\ \epsilon>0$, $\exists\ 0<c_\epsilon=(1-\epsilon)\cdot(e+\epsilon)^{\frac{(-\frac{e}{2}-\epsilon)\cdot(1+\epsilon)}{e\cdot(1-\epsilon)}}<1$ *and* $y_\epsilon>0$, *such that* $\frac{c_\epsilon\cdot(\frac{ex}{y})^y}{\sqrt{2\pi y}}\leq\binom{x}{y}\leq\frac{e^\epsilon\cdot(\frac{ex}{y})^y}{\sqrt{2\pi y}}$ *for* $x\geq y^2$ *and* $y\geq y_\epsilon$. $\square$

Next we prove Proposition 2.3, which gives the lower bound, $c\cdot\log m$, of the average information entropy of hypergeometric distributions $HG(m,n)$.

Proposition 2.3 *Let* $HG(m,n)=-n^{-1}\cdot\sum_{j\in[n]}\sum_{i\in[m]}\frac{\binom{j-1}{i-1}\cdot\binom{n-j}{m-i}}{\binom{n-1}{m-1}}\cdot\log\frac{\binom{j-1}{i-1}\cdot\binom{n-j}{m-i}}{\binom{n-1}{m-1}}$. *Then* $\exists$ $0<c<1$ *and* $m_c>0$, *such that* $HG(m,n)\geq c\cdot\log m$ *for* $n\geq m^2$ *and* $m\geq m_c$. *Specifically, we show that* c *is a constant.*

Proof. *Let* $0<\sigma<\frac{\sqrt{5}}{2}-1$. *For each* $j\in[\frac{n}{2}-\sigma\cdot n,\frac{n}{2}+\sigma\cdot n]$, *we associate* j *with* $m^{\frac{1}{2}}$ *many* i *where* $i\in[\frac{j}{n}\cdot m+m^{\frac{1}{2}},\frac{j}{n}\cdot m+2m^{\frac{1}{2}}]$. *First we derive the conditions such that Lemma 2.2 can be applied to* $\binom{j-1}{i-1}$, $\binom{n-j}{m-i}$, *and* $\binom{n-1}{m-1}$ *where* $j\in[\frac{n}{2}-\sigma\cdot n,\frac{n}{2}+\sigma\cdot n]$ *and* $i\in[\frac{j}{n}\cdot m+m^{\frac{1}{2}},\frac{j}{n}\cdot m+2m^{\frac{1}{2}}]$.

(1) $\forall\ \epsilon>0$, $\exists\ 0<c_\epsilon<1$ *and* $y_\epsilon>0$ *such that* $\frac{c_\epsilon\cdot(\frac{e(j-1)}{i-1})^{i-1}}{\sqrt{2\pi(i-1)}}\leq\binom{j-1}{i-1}\leq\frac{e^\epsilon\cdot(\frac{e(j-1)}{i-1})^{i-1}}{\sqrt{2\pi(i-1)}}$ *for* $j-1\geq(i-1)^2$ *and* $i-1\geq y_\epsilon(*_1)$ *according to Lemma 2.2.*

(2) $\forall\ \epsilon>0$, $\exists\ 0<c_\epsilon<1$ *and* $y_\epsilon>0$ *such that* $\frac{c_\epsilon\cdot(\frac{e(n-j)}{m-i})^{m-i}}{\sqrt{2\pi(m-i)}}\leq\binom{n-j}{m-i}\leq\frac{e^\epsilon\cdot(\frac{e(n-j)}{m-i})^{m-i}}{\sqrt{2\pi(m-i)}}$ *for* $n-j\geq(m-i)^2$ *and* $m-i\geq y_\epsilon(*_2)$ *according to Lemma 2.2.*

(3) $\forall\ \epsilon>0$, $\exists\ 0<c_\epsilon<1$ *and* $y_\epsilon>0$ *such that* $\frac{c_\epsilon\cdot(\frac{e(n-1)}{m-1})^{m-1}}{\sqrt{2\pi(m-1)}}\leq\binom{n-1}{m-1}\leq\frac{e^\epsilon\cdot(\frac{e(n-1)}{m-1})^{m-1}}{\sqrt{2\pi(m-1)}}$ *for* $n-1\geq(m-1)^2$ *and* $m-1\geq y_\epsilon(*_3)$ *according to Lemma 2.2.*

Note that $(\frac{1}{2}-\sigma)\cdot n\leq j\leq(\frac{1}{2}+\sigma)\cdot n$ *and* $(\frac{1}{2}-\sigma)\cdot m+m^{\frac{1}{2}}\leq\frac{j}{n}\cdot m+m^{\frac{1}{2}}\leq i\leq\frac{j}{n}\cdot m+2m^{\frac{1}{2}}\leq(\frac{1}{2}+\sigma)\cdot m+2m^{\frac{1}{2}}$. *Then* $(\frac{1}{2}-\sigma)\cdot n\leq n-j\leq(\frac{1}{2}+\sigma)\cdot n$ *and* $(\frac{1}{2}-\sigma)\cdot m-2m^{\frac{1}{2}}\leq m-i\leq(\frac{1}{2}+\sigma)\cdot m-m^{\frac{1}{2}}$.

Consider condition $(*_1)$. $j-1\geq(\frac{1}{2}-\sigma)\cdot n-1$, $i-1\leq(\frac{1}{2}+\sigma)\cdot m+2m^{\frac{1}{2}}-1$. *Since* $0<\sigma<\frac{\sqrt{5}}{2}-1$, $(\frac{1}{2}-\sigma)\cdot n\geq(\frac{1}{2}-\sigma)\cdot m^2>(\frac{1}{2}+\sigma)^2\cdot m^2$ *for* $n\geq m^2$. *Therefore* $\exists\ m_{\sigma,\epsilon,1}>0$ *such that* $j-1\geq(i-1)^2$ *and* $i-1\geq y_\epsilon$ *for* $n\geq m^2$ *and* $m\geq m_{\sigma,\epsilon,1}$.

Consider condition $(*_2)$. $n-j \geq (\frac{1}{2}-\sigma)\cdot n$, $m-i \leq (\frac{1}{2}+\sigma)\cdot m - m^{\frac{1}{2}}$. *Since* $0<\sigma<\frac{\sqrt{5}}{2}-1$, $(\frac{1}{2}-\sigma)\cdot n \geq (\frac{1}{2}-\sigma)\cdot m^2 > (\frac{1}{2}+\sigma)^2\cdot m^2$ *for* $n \geq m^2$. *Therefore* $\exists\ m_{\sigma,\epsilon,2}>0$ *such that* $n-j \geq (m-i)^2$ *and* $m-i \geq y_\epsilon$ *for* $n \geq m^2$ *and* $m \geq m_{\sigma,\epsilon,2}$.

Consider condition $(*_3)$. $n-1 \geq m^2-1 \geq m^2-2m+1=(m-1)^2$. *Therefore* $\exists\ m_{\epsilon,3}>0$ *such that* $n-1 \geq (m-1)^2$ *and* $m-1 \geq y_\epsilon$ *for* $n \geq m^2$ *and* $m \geq m_{\epsilon,3}$.

Therefore $\exists\ m_{\sigma,\epsilon,4}>0$ *such that the estimation of Lemma 2.2 can be applied to* $\binom{j-1}{i-1}$, $\binom{n-j}{m-i}$, *and* $\binom{n-1}{m-1}$ *for* $j \in [\frac{n}{2}-\sigma\cdot n, \frac{n}{2}+\sigma\cdot n]$, $i \in [\frac{j}{n}\cdot m + m^{\frac{1}{2}}, \frac{j}{n}\cdot m + 2m^{\frac{1}{2}}]$, $n \geq m^2$ *and* $m \geq m_{\sigma,\epsilon,4}$. *Thus* $\frac{c_\epsilon^2}{e^\epsilon}\cdot\frac{\frac{(\frac{e(j-1)}{i-1})^{i-1}}{\sqrt{2\pi(i-1)}}\cdot\frac{(\frac{e(n-j)}{m-i})^{m-i}}{\sqrt{2\pi(m-i)}}}{\frac{(\frac{e(n-1)}{m-1})^{m-1}}{\sqrt{2\pi(m-1)}}} \leq \frac{\binom{j-1}{i-1}\cdot\binom{n-j}{m-i}}{\binom{n-1}{m-1}} \leq \frac{e^{2\epsilon}}{c_\epsilon}\frac{\frac{(\frac{e(j-1)}{i-1})^{i-1}}{\sqrt{2\pi(i-1)}}\cdot\frac{(\frac{e(n-j)}{m-i})^{m-i}}{\sqrt{2\pi(m-i)}}}{\frac{(\frac{e(n-1)}{m-1})^{m-1}}{\sqrt{2\pi(m-1)}}}$. *Consider the term* $\frac{\frac{(\frac{e(j-1)}{i-1})^{i-1}}{\sqrt{2\pi(i-1)}}\cdot\frac{(\frac{e(n-j)}{m-i})^{m-i}}{\sqrt{2\pi(m-i)}}}{\frac{(\frac{e(n-1)}{m-1})^{m-1}}{\sqrt{2\pi(m-1)}}} = \frac{1}{\sqrt{2\pi}}\cdot\sqrt{\frac{m-1}{(i-1)\cdot(m-i)}}\cdot(\frac{j-1}{n-1}\cdot\frac{m-1}{i-1})^{i-1}\cdot(\frac{n-j}{n-1}\cdot\frac{m-1}{m-i})^{m-i}$. *Let* $T_1=\sqrt{\frac{m-1}{(i-1)\cdot(m-i)}}$, $T_2=T_{21}\cdot T_{22}$ *where* $T_{21}=(\frac{j-1}{n-1}\cdot\frac{m-1}{i-1})^{i-1}$ *and* $T_{22}=(\frac{n-j}{n-1}\cdot\frac{m-1}{m-i})^{m-i}$. *Then* $\frac{\frac{(\frac{e(j-1)}{i-1})^{i-1}}{\sqrt{2\pi(i-1)}}\cdot\frac{(\frac{e(n-j)}{m-i})^{m-i}}{\sqrt{2\pi(m-i)}}}{\frac{(\frac{e(n-1)}{m-1})^{m-1}}{\sqrt{2\pi(m-1)}}} = \frac{1}{\sqrt{2\pi}}\cdot T_1\cdot T_2$. *Next we estimate* T_1 *and* T_2, *respectively.*

Consider $T_1 = \sqrt{\frac{m-1}{(i-1)\cdot(m-i)}}$, $\exists\ m_{\sigma,\epsilon,5}>0$ *such that* $\frac{1}{(\frac{1}{2}+\sigma+\epsilon)\cdot\sqrt{m}} = \frac{1}{\sqrt{(\frac{1}{2}+\sigma+\epsilon)^2\cdot m}} < \sqrt{\frac{m-1}{((\frac{1}{2}+\sigma)\cdot m+2m^{\frac{1}{2}}-1)\cdot((\frac{1}{2}+\sigma)\cdot m-m^{\frac{1}{2}})}} \leq T_1 = \sqrt{\frac{m-1}{(i-1)\cdot(m-i)}} \leq \sqrt{\frac{m-1}{((\frac{1}{2}-\sigma)\cdot m+m^{\frac{1}{2}}-1)\cdot((\frac{1}{2}-\sigma)\cdot m-2m^{\frac{1}{2}})}} < \frac{1}{\sqrt{(\frac{1}{2}-\sigma-\epsilon)^2\cdot m}} = \frac{1}{(\frac{1}{2}-\sigma-\epsilon)\cdot\sqrt{m}}$ *for* $m \geq m_{\sigma,\epsilon,5}$.

Consider $T_2 = T_{21}\cdot T_{22}$ *where* $T_{21}=(\frac{j-1}{n-1}\cdot\frac{m-1}{i-1})^{i-1}$ *and* $T_{22}=(\frac{n-j}{n-1}\cdot\frac{m-1}{m-i})^{m-i}$. *Denote* $j=\frac{n}{2}+u$ *where* $-\sigma\cdot n \leq u \leq \sigma\cdot n$ *and* $i=\frac{j}{n}\cdot m+v=(\frac{1}{2}+\frac{u}{n})\cdot m+v$ *where* $m^{\frac{1}{2}} \leq v \leq 2m^{\frac{1}{2}}$. *Then* $T_{21}=(\frac{j-1}{n-1}\cdot\frac{m-1}{i-1})^{i-1} = (\frac{\frac{n}{2}+u-1}{n-1}\cdot\frac{m-1}{(\frac{1}{2}+\frac{u}{n})\cdot m+v-1})^{(\frac{1}{2}+\frac{u}{n})\cdot m+v-1} = ((1+\frac{1-\frac{1}{\frac{1}{2}+\frac{u}{n}}}{n-1})\cdot(1-\frac{\frac{v-1}{\frac{1}{2}+\frac{u}{n}}+1}{m+\frac{v-1}{\frac{1}{2}+\frac{u}{n}}}))^{(\frac{1}{2}+\frac{u}{n})\cdot m+v-1} = (1+\frac{1-\frac{1}{\frac{1}{2}+\frac{u}{n}}}{n-1})^{(\frac{1}{2}+\frac{u}{n})\cdot m+v-1}\cdot(1-\frac{\frac{v-1}{\frac{1}{2}+\frac{u}{n}}+1}{m+\frac{v-1}{\frac{1}{2}+\frac{u}{n}}})^{(\frac{1}{2}+\frac{u}{n})\cdot m+v-1}$. $\exists\ m_{\sigma,\epsilon,6}>0$ *such that* $1-\epsilon < (1+\frac{1-\frac{1}{\frac{1}{2}+\frac{u}{n}}}{n-1})^{(\frac{1}{2}+\frac{u}{n})\cdot m+v-1} < 1$ *for* $n \geq m^2$ *and* $m \geq m_{\sigma,\epsilon,6}$. $(1-\frac{\frac{v-1}{\frac{1}{2}+\frac{u}{n}}+1}{m+\frac{v-1}{\frac{1}{2}+\frac{u}{n}}})^{(\frac{1}{2}+\frac{u}{n})\cdot m+v-1} = (e-\frac{c_1\cdot(\frac{v-1}{\frac{1}{2}+\frac{u}{n}}+1)}{m+\frac{v-1}{\frac{1}{2}+\frac{u}{n}}})^{-\frac{\frac{v-1}{\frac{1}{2}+\frac{u}{n}}+1}{m+\frac{v-1}{\frac{1}{2}+\frac{u}{n}}}\cdot[(\frac{1}{2}+\frac{u}{n})\cdot m+v-1]} = (e-\frac{c_1\cdot(\frac{v-1}{\frac{1}{2}+\frac{u}{n}}+1)}{m+\frac{v-1}{\frac{1}{2}+\frac{u}{n}}})^{-v+\frac{1}{2}-\frac{u}{n}}$ *for* $|-\frac{m+\frac{v-1}{\frac{1}{2}+\frac{u}{n}}}{\frac{v-1}{\frac{1}{2}+\frac{u}{n}}+1}| \geq x_\epsilon$. *Hence* $\exists\ m_{\sigma,\epsilon,7}>0$ *such that* $(e+\frac{2(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}(\frac{1}{2}-\sigma)})^{-v+\frac{1}{2}-\frac{u}{n}} \leq (e-\frac{c_1\cdot\frac{v}{\frac{1}{2}+\frac{u}{n}}}{m})^{-v+\frac{1}{2}-u} \leq (e-\frac{c_1\cdot\frac{v-1}{\frac{1}{2}+\frac{u}{n}}}{m})^{-v+\frac{1}{2}-\frac{u}{n}} \leq (1-\frac{\frac{v-1}{\frac{1}{2}+\frac{u}{n}}+1}{m+\frac{v-1}{\frac{1}{2}+\frac{u}{n}}})^{(\frac{1}{2}+\frac{u}{n})\cdot m+v-1} = (e-\frac{c_1\cdot(\frac{v-1}{\frac{1}{2}+\frac{u}{n}}+1)}{m+\frac{v-1}{\frac{1}{2}+\frac{u}{n}}})^{-v+\frac{1}{2}-\frac{u}{n}} \leq e^{-v+\frac{1}{2}-\frac{u}{n}}$ *for* $n \geq m^2$ *and* $m \geq m_{\sigma,\epsilon,7}$.

Thus $(1-\epsilon)\cdot(e+\frac{2(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}(\frac{1}{2}-\sigma)})^{-v+\frac{1}{2}-\frac{u}{n}} \le T_{21} \le e^{-v+\frac{1}{2}-\frac{u}{n}}$ *for* $n \ge m^2$ *and* $m \ge \max\{m_{\sigma,\epsilon,6}, m_{\sigma,\epsilon,7}\}$

Similarly $T_{22} = (\frac{n-j}{n-1}\cdot\frac{m-1}{m-i})^{m-i} = (\frac{\frac{n}{2}-u}{n-1}\cdot\frac{m-1}{(\frac{1}{2}-\frac{u}{n})\cdot m-v})^{(\frac{1}{2}-\frac{u}{n})\cdot m-v} = ((1+\frac{1}{n-1})\cdot(1+\frac{\frac{v}{\frac{1}{2}-\frac{u}{n}}-1}{m-\frac{v}{\frac{1}{2}-\frac{u}{n}}}))^{(\frac{1}{2}-\frac{u}{n})\cdot m-v} = (1+\frac{1}{n-1})^{(\frac{1}{2}-\frac{u}{n})\cdot m-v}\cdot(1+\frac{\frac{v}{\frac{1}{2}-\frac{u}{n}}-1}{m-\frac{v}{\frac{1}{2}-\frac{u}{n}}})^{(\frac{1}{2}-\frac{u}{n})\cdot m-v}$. $\exists\ m_{\sigma,\epsilon,8} > 0$ *such that* $1 < (1+\frac{1}{n-1})^{(\frac{1}{2}-\frac{u}{n})\cdot m-v} < 1+\epsilon$ *for* $n \ge m^2$ *and* $m \ge m_{\sigma,\epsilon,8}$. $(1+\frac{\frac{v}{\frac{1}{2}-\frac{u}{n}}-1}{m-\frac{v}{\frac{1}{2}-\frac{u}{n}}})^{(\frac{1}{2}-\frac{u}{n})\cdot m-v} = (e+\frac{c_1\cdot(\frac{v}{\frac{1}{2}-\frac{u}{n}}-1)}{m-\frac{v}{\frac{1}{2}-\frac{u}{n}}})^{\frac{\frac{v}{\frac{1}{2}-\frac{u}{n}}-1}{m-\frac{v}{\frac{1}{2}-\frac{u}{n}}}\cdot[(\frac{1}{2}-\frac{u}{n})\cdot m-v]} = (e+\frac{c_1\cdot(\frac{v}{\frac{1}{2}-\frac{u}{n}}-1)}{m-\frac{v}{\frac{1}{2}-\frac{u}{n}}})^{v-\frac{1}{2}+\frac{u}{n}}$ *for* $|\frac{m-\frac{v}{\frac{1}{2}-\frac{u}{n}}}{\frac{v}{\frac{1}{2}-\frac{u}{n}}-1}| \ge x_\epsilon$. *Hence* $\exists\ m_{\sigma,\epsilon,9} > 0$ *such that* $(e-\frac{2(1+\epsilon)\cdot(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)})^{v-\frac{1}{2}+\frac{u}{n}} \le (e+\frac{c_1\cdot(1+\epsilon)\cdot\frac{v}{\frac{1}{2}-\frac{u}{n}}}{m})^{v-\frac{1}{2}+\frac{u}{n}} \le (e+\frac{c_1\cdot(1+\epsilon)\cdot(\frac{v}{\frac{1}{2}-\frac{u}{n}}-1)}{m})^{v-\frac{1}{2}+\frac{u}{n}} \le (1+\frac{\frac{v}{\frac{1}{2}-\frac{u}{n}}-1}{m-\frac{v}{\frac{1}{2}-\frac{u}{n}}})^{(\frac{1}{2}-\frac{u}{n})\cdot m-v} = (e+\frac{c_1\cdot(\frac{v}{\frac{1}{2}-\frac{u}{n}}-1)}{m-\frac{v}{\frac{1}{2}-\frac{u}{n}}})^{v-\frac{1}{2}+\frac{u}{n}} \le e^{v-\frac{1}{2}+\frac{u}{n}}$ *for* $n \ge m^2$ *and* $m \ge m_{\sigma,\epsilon,9}$. *Thus* $(e-\frac{2(1+\epsilon)\cdot(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)})^{v-\frac{1}{2}+\frac{u}{n}} \le T_{22} \le (1+\epsilon)\cdot e^{v-\frac{1}{2}+\frac{u}{n}}$ *for* $n \ge m^2$ *and* $m \ge \max\{m_{\sigma,\epsilon,8}, m_{\sigma,\epsilon,9}\}$.

Hence $(1-\epsilon)\cdot(e+\frac{2(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)})^{-v+\frac{1}{2}-\frac{u}{n}}\cdot(e-\frac{2(1+\epsilon)\cdot(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)})^{v-\frac{1}{2}+\frac{u}{n}} \le T_2 = T_{21}\cdot T_{22} \le e^{-v+\frac{1}{2}-\frac{u}{n}}\cdot(1+\epsilon)\cdot e^{v-\frac{1}{2}+\frac{u}{n}} = 1+\epsilon$. *Note that* $(e+\frac{2(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)})^{-v+\frac{1}{2}-\frac{u}{n}}\cdot(e-\frac{2(1+\epsilon)\cdot(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)})^{v-\frac{1}{2}+\frac{u}{n}} = (e+\frac{2(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)})^{\frac{1}{2}-\frac{u}{n}}\cdot(e-\frac{2(1+\epsilon)\cdot(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)})^{-\frac{1}{2}+\frac{u}{n}}\cdot(\frac{e-\frac{2(1+\epsilon)\cdot(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)}}{e+\frac{2(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)}})^{v}$. $\exists\ m_{\sigma,\epsilon,10} > 0$ *such that* $e^{\frac{1}{2}-\sigma} \le (e+\frac{2(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)})^{\frac{1}{2}-\frac{u}{n}} \le (e+\epsilon)^{\frac{1}{2}+\sigma}$ *for* $m \ge m_{\sigma,\epsilon,10}$. $\exists\ m_{\sigma,\epsilon,11} > 0$ *such that* $e^{-\frac{1}{2}-\sigma} \le (e-\frac{2(1+\epsilon)\cdot(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)})^{-\frac{1}{2}+\frac{u}{n}} \le (e-\epsilon)^{-\frac{1}{2}+\sigma}$ *for* $m \ge m_{\sigma,\epsilon,11}$. $\exists\ m_{\sigma,\epsilon,12} > 0$ *such that* $(\frac{e-\frac{2(1+\epsilon)\cdot(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)}}{e+\frac{2(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)}})^{v} \ge (\frac{e-\frac{2(1+\epsilon)\cdot(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)}}{e+\frac{2(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)}})^{2m^{\frac{1}{2}}} = (1-\frac{(4+2\epsilon)\cdot(\frac{e}{2}+\epsilon)}{e\cdot(\frac{1}{2}-\sigma)\cdot m^{\frac{1}{2}}+2(\frac{e}{2}+\epsilon)})^{2m^{\frac{1}{2}}} = (e-\frac{c_1\cdot(4+2\epsilon)\cdot(\frac{e}{2}+\epsilon)}{e\cdot(\frac{1}{2}-\sigma)\cdot m^{\frac{1}{2}}+2(\frac{e}{2}+\epsilon)})^{-\frac{(4+2\epsilon)\cdot(\frac{e}{2}+\epsilon)}{e\cdot(\frac{1}{2}-\sigma)\cdot m^{\frac{1}{2}}+2(\frac{e}{2}+\epsilon)}\cdot 2m^{\frac{1}{2}}} = (e-\frac{c_1\cdot(4+2\epsilon)\cdot(\frac{e}{2}+\epsilon)}{e\cdot(\frac{1}{2}-\sigma)\cdot m^{\frac{1}{2}}+2(\frac{e}{2}+\epsilon)})^{-\frac{(8+4\epsilon)\cdot(\frac{e}{2}+\epsilon)\cdot m^{\frac{1}{2}}}{e\cdot(\frac{1}{2}-\sigma)\cdot m^{\frac{1}{2}}+2(\frac{e}{2}+\epsilon)}} \ge (e+\epsilon)^{-\frac{(8+4\epsilon)\cdot(\frac{e}{2}+\epsilon)}{e\cdot(\frac{1}{2}-\sigma)}}$ *for* $m \ge m_{\sigma,\epsilon,12}$. *Hence* $(e+\epsilon)^{-\frac{(8+4\epsilon)\cdot(\frac{e}{2}+\epsilon)}{e\cdot(\frac{1}{2}-\sigma)}} \le (\frac{e-\frac{2(1+\epsilon)\cdot(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)}}{e+\frac{2(\frac{e}{2}+\epsilon)}{m^{\frac{1}{2}}\cdot(\frac{1}{2}-\sigma)}})^{v} < 1$ *for* $m \ge m_{\sigma,\epsilon,12}$. *Therefore* $(1-\epsilon)\cdot e^{-2\sigma}\cdot(e+\epsilon)^{-\frac{(8+4\epsilon)\cdot(\frac{e}{2}+\epsilon)}{e\cdot(\frac{1}{2}-\sigma)}} = (1-\epsilon)\cdot e^{\frac{1}{2}-\sigma}\cdot e^{-\frac{1}{2}-\sigma}\cdot(e+\epsilon)^{-\frac{(8+4\epsilon)\cdot(\frac{e}{2}+\epsilon)}{e\cdot(\frac{1}{2}-\sigma)}} \le T_2 = T_{21}\cdot T_{22} \le 1+\epsilon$.

Hence $\forall\ \epsilon > 0$, $0 < \sigma < \frac{\sqrt{5}}{2}-1$, $\exists\ m_{\sigma,\epsilon,13} > 0$ *such that* $\frac{c_\epsilon^2}{\sqrt{2\pi}\cdot e^\epsilon}\cdot\frac{1}{(\frac{1}{2}+\sigma+\epsilon)\cdot\sqrt{m}}\cdot(1-\epsilon)\cdot e^{-2\sigma}\cdot(e+$

$\epsilon)^{-\frac{(8+4\epsilon)\cdot(\frac{e}{2}+\epsilon)}{e\cdot(\frac{1}{2}-\sigma)}} \leq \frac{\binom{j-1}{i-1}\cdot\binom{n-j}{m-i}}{\binom{n-1}{m-1}} \leq \frac{e^{2\epsilon}}{\sqrt{2\pi}\cdot c_\epsilon}\cdot\frac{1}{(\frac{1}{2}-\sigma-\epsilon)\cdot\sqrt{m}}\cdot(1+\epsilon)$ *for* $j \in [\frac{n}{2}-\sigma\cdot n, \frac{n}{2}+\sigma\cdot n]$, $i \in [\frac{j}{n}\cdot m+m^{\frac{1}{2}}, \frac{j}{n}\cdot m+2m^{\frac{1}{2}}]$, $n \geq m^2$, *and* $m \geq m_{\sigma,\epsilon,13}$. *Let* $c_{\epsilon,\sigma,1} = \frac{c_\epsilon^2}{\sqrt{2\pi}\cdot e^\epsilon}\cdot\frac{1}{(\frac{1}{2}+\sigma+\epsilon)}\cdot(1-\epsilon)\cdot e^{-2\sigma}\cdot(e+\epsilon)^{-\frac{(8+4\epsilon)\cdot(\frac{e}{2}+\epsilon)}{e\cdot(\frac{1}{2}-\sigma)}}$ *and* $c_{\epsilon,\sigma,2} = \frac{e^{2\epsilon}}{\sqrt{2\pi}\cdot c_\epsilon}\cdot\frac{1}{(\frac{1}{2}-\sigma-\epsilon)}\cdot(1+\epsilon)$. *Then* $c_{\epsilon,\sigma,1}\cdot m^{-\frac{1}{2}} \leq \frac{\binom{j-1}{i-1}\cdot\binom{n-j}{m-i}}{\binom{n-1}{m-1}} \leq c_{\epsilon,\sigma,2}\cdot m^{-\frac{1}{2}}$ *for* $j \in [\frac{n}{2}-\sigma\cdot n, \frac{n}{2}+\sigma\cdot n]$, $i \in [\frac{j}{n}\cdot m+m^{\frac{1}{2}}, \frac{j}{n}\cdot m+2m^{\frac{1}{2}}]$, $n \geq m^2$, *and* $m \geq m_{\sigma,\epsilon,13}$. *Thus* $HG(m,n) = -n^{-1}\cdot\sum_{j\in[n]}\sum_{i\in[m]} \frac{\binom{j-1}{i-1}\cdot\binom{n-j}{m-i}}{\binom{n-1}{m-1}}\cdot\log\frac{\binom{j-1}{i-1}\cdot\binom{n-j}{m-i}}{\binom{n-1}{m-1}} \geq (-n^{-1})\cdot(2\sigma\cdot n)\cdot m^{\frac{1}{2}}\cdot(c_{\epsilon,\sigma,1}\cdot m^{-\frac{1}{2}})\cdot\log(c_{\epsilon,\sigma,2}\cdot m^{-\frac{1}{2}}) = \sigma\cdot c_{\epsilon,\epsilon,1}\cdot\log m - 2\sigma\cdot c_{\epsilon,\sigma,1}\cdot\log c_{\epsilon,\sigma,2} = (1-\frac{2\log c_{\epsilon,\sigma,2}}{\log m})\cdot(\sigma\cdot c_{\epsilon,\sigma,1}\cdot\log m) = (1-o(1))\cdot(\sigma\cdot c_{\epsilon,\sigma,1}\cdot\log m)$. *Hence* $\forall\ \epsilon > 0$ *and* $0 < \sigma < \frac{\sqrt{5}}{2}-1$, $\exists\ 0 < c < 1$ *and* $m_c > 0$, *such that* $HG(m,n) \geq c\cdot\log m$ *for* $n \geq m^2$ *and* $m \geq m_c$. □

3 Information Leaks in the $SE_{m,n}$ Scheme

Let $[m] = \{i \mid 1 \leq i \leq m\}$ be the domain of plain texts and $[n] = \{j \mid 1 \leq j \leq n\}$ be the range of cipher texts, where $m \leq n$. Suppose that $SE_{m,n} = (K_{m,n}, E_{m,n}, D_{m,n})$ is an OPE scheme where $K_{m,n} : \{0,1\}^* \to \{0,1\}^*$ is the key generation algorithm, $E_{m,n} : [m] \times \{0,1\}^* \to [n]$ is the encryption algorithm, and $D_{m,n} : [n] \times \{0,1\}^* \to [m]$ is the decryption algorithm satisfying $D_{m,n}(E_{m,n}(x,k),k) = x$ for all $x \in [m]$ and $E_{m,n}(x,k) \leq E_{m,n}(x',k)$ for all $x \leq x'$. In [8], the security of $SE_{m,n}$ is defined based on the ideal model and real model.

Definition 3.1 *The ideal model: Let the set* $SIF_{m,n} = \{f \mid f : [m] \to [n]$ *is a strictly increasing function*$\}$. *Let* $R_{m,n}$ *be the function selected uniformly randomly from* $SIF_{m,n}$. *For plain text* x, *the cipher text* $y = R_{m,n}(x)$.

The real model: For plain text x, *the cipher text* $y = E_{m,n}(x,k)$.

$SE_{m,n}$ *is secure if the cipher text* $R_{m,n}(x)$ *and the cipher text* $E_{m,n}(x,k)$ *are computationally indistinguishable.* □

Without loss of generality, let $SE_{m,n} = (K_{m,n}, E_{m,n}, D_{m,n})$ denote the OPE scheme constructed in [8]. As proven in [8], $SE_{m,n}$ is secure because $R_{m,n}(x)$ and $E_{m,n}(x,k)$ are computationally indistinguishable.

Next, we assume that the plain texts are uniformly distributed and analyze the security of $SE_{m,n}$. Since the encryption results of $R_{m,n}(x)$ and $E_{m,n}(x,k)$ are computationally indistinguishable, it suffices to analyze the security of $R_{m,n}$. In [8], the authors proved that no OPE algorithm can achieve indistinguishable security under chosen plain text attacks. Here, we analyze the security in a differently aspect. We first show that the plain text encrypted using an OPE algorithm may have information leaks under chosen plain text attacks. Such analysis can help us develop the method to estimate an upper bound on the probability for the adversary to recover the plain text encrypted by $R_{m,n}$ under chosen plain text attacks.

The chosen plain text attack can be divided into two phases. In the first phase, the adversary chooses plain texts x_i from $[m]$, denoted as $x_i \xleftarrow{A} [m]$, and gets the plain text and cipher text pairs $(x_i, OPE(x_i, k))$, $1 \le i \le i_0$. In the second phase, the adversary is given a cipher text $OPE(x, k)$, where x is selected uniformly randomly from $[m]$, denoted as $x \xleftarrow{\$} [m]$. The security of the OPE algorithm under the chosen plain text attack depends on how much information of x the adversary can retrieve.

Lemma 3.1 shows a lower bound of how much information of a plain text the adversary can retrieve with chosen plain text attacks.

Lemma 3.1 *For an OPE algorithm, if an adversary A can choose i_0 plain texts $x_i, 1 \le i \le i_0$, then given a cipher text $OPE(x, k)$, A can retrieve at least $\log(i_0 + 1)$ bits of information of x.*

Proof. *Suppose that A chooses the plain texts $x_i = \frac{i \cdot m}{2^{\log(i_0+1)}}$ and gets $(x_i, OPE(x_i, k))$, $1 \le i \le i_0$. For the cipher text $OPE(x, k)$, $\exists$ i' such that $OPE(\frac{(i'-1) \cdot m}{2^{\log(i_0+1)}}, k) \le OPE(x, k) < OPE(\frac{i' \cdot m}{2^{\log(i_0+1)}}, k)$. Then A can derive that $\frac{(i'-1) \cdot m}{2^{\log(i_0+1)}} \le x < \frac{i' \cdot m}{2^{\log(i_0+1)}}$. The information entropy of the event "$\frac{(i'-1) \cdot m}{2^{\log(i_0+1)}} \le x < \frac{i' \cdot m}{2^{\log(i_0+1)}}$" is $\log(\frac{i' \cdot m}{2^{\log(i_0+1)}} - \frac{(i'-1) \cdot m}{2^{\log(i_0+1)}}) = \log \frac{m}{2^{\log(i_0+1)}} = \log m - \log(i_0 + 1)$. Hence the adversary can retrieve at least $\log m - (\log m - \log(i_0 + 1)) = \log(i_0 + 1)$ bits of information of x.* $\square$

Here is an example to illustrate Lemma 3.1. Suppose that $[m] = [\overbrace{0 \cdots 0}^{l-1} 1, \overbrace{1 \cdots 1}^{l}]$ where $\overbrace{0 \cdots 0}^{l-1} 1$ and $\overbrace{1 \cdots 1}^{l}$ are binary numbers. The adversary chooses the plain text $1 \overbrace{0 \cdots 0}^{l-1}$ and gets the plain text and cipher text pair $(1 \overbrace{0 \cdots 0}^{l-1}, OPE(1 \overbrace{0 \cdots 0}^{l-1}, k))$. Suppose that the adversary is given a cipher text $OPE(x, k)$. The adversary compares $OPE(x, k)$ with $OPE(1 \overbrace{0 \cdots 0}^{l-1}, k)$. If $OPE(x, k) > OPE(1 \overbrace{0 \cdots 0}^{l-1}, k)$, then x is in $(1 \overbrace{0 \cdots 0}^{l-1}, \overbrace{1 \cdots 1}^{l}]$ and, hence, the most significant bit of $x = 1$. If $OPE(x, k) < OPE(1 \overbrace{0 \cdots 0}^{l-1}, k)$, then x is in $[\overbrace{0 \cdots 0}^{l-1} 1, 1 \overbrace{0 \cdots 0}^{l-1})$ and, hence, the most significant bit of $x = 0$. Thus the adversary can retrieve at least 1 bit of information of the plain text x.

Even if there is no chosen plain text attacks, i.e., $i_0 = 0$, there are still information leaks. Assume that the adversary is given the cipher text $OPE(x, k)$, the adversary can guess x. For example, if $OPE(x, k) = 1$ or n, then x must be 1 or m. Also, for the general situation, the adversary can guess x from a probability distribution. Since this probability distribution is not a uniform distribution, the corresponding information entropy does not achieve the maximal possible value $\log m$. Therefore, there are information leaks even if $i_0 = 0$.

4 Security Analysis for the $SE_{m,n}$ Scheme

In this section, we derive an upper bound on the probability for the adversary to recover a plain text x under chosen plain text attacks. Let z denote the information entropy of the part of x that remains secret from the adversary under chosen plain text attacks. Then, according to information theory, the probability for the adversary to recover x is at most 2^{-z}. Thus, we need to consider the problem regarding how much information of a plain text encrypted by $R_{m,n}$ can remain secret from the adversary under chosen plain text attacks.

First, we consider the special case when the adversary has no knowledge of any plain text and cipher text pairs. In this case, our task is to show that the information entropy of the plain text encrypted by $R_{m,n}$ is the average information entropy of hypergeometric distribution, which has been derived in Proposition 2.3. We give this proof in Subsection 4.1.

In Section 4.2, we consider the case that the adversary does have the knowledge of m^a plain text and cipher text pairs, for some a, $0 < a < 1$. Based on the discussions in Section 3 and the proof of the special case, we can derive how much information of a plain text encrypted by $R_{m,n}$ can remain secret from the adversary under chosen plain text attacks. Accordingly, we derive the upper bound on the probability for the adversary to recover a plain text x under chosen plain text attacks.

4.1 No Chosen Plain Text Attacks

Suppose that the adversary is given cipher text $R_{m,n}(x)$ with no plain text and cipher text pairs. In Proposition 4.1, we prove that the distribution of the plain text x follows the hypergeometric distribution from the adversary's point of view. More specifically, we prove that the upper bound on the probability for the adversary to recover x is $2^{-c\cdot\log m}$, where $c\cdot\log m$ has been shown to be the lower bound of the average information entropy of hypergeometric distributions (Section 2).

Proposition 4.1 *Let A be an adversary. $\exists\ 0 < c < 1$ and $m_c > 0$, s.t. $Pr[A(R_{m,n}(x)) = x : R_{m,n} \xleftarrow{\$} SIF_{m,n}; x \xleftarrow{\$} [m]] \leq 2^{-c\cdot\log m}$ for $n \geq m^2$ and $m \geq m_c$.*

Proof. *For $j \in [n]$, let E_j be the event that the adversary views j. Let $SIF_{m,n}(j) = \{f \in SIF_{m,n} \mid \exists\ i \in [m]$ such that $f(i) = j\}$. Then $|SIF_{m,n}(j)| = \sum_{i\in[m]} |f \in SIF_{m,n} | f(i) = j| = \sum_{i\in[m]} \binom{j-1}{i-1} \cdot \binom{n-j}{m-i} = \binom{n-1}{m-1}$. Hence $Pr[E_j] = \sum_{g\in SIF_{m,n}} Pr[E_j|g] \cdot Pr[g] = \frac{\sum_{g\in SIF_{m,n}} Pr[E_j|g]}{\binom{n}{m}} = \frac{\sum_{g\in SIF_{m,n}(j)} Pr[E_j|g]}{\binom{n}{m}} = \frac{\sum_{g\in SIF_{m,n}(j)} \frac{1}{m}}{\binom{n}{m}} = \frac{|SIF_{m,n}(j)|}{m\cdot\binom{n}{m}} = \frac{\binom{n-1}{m-1}}{m\cdot\binom{n}{m}} = \frac{1}{n}$.*

For $i \in [m]$ and $j \in [n]$, let "$i \to j$" be the event that j is mapped from i. Then "$i \to j|E_j$" is the event that j is mapped from i given the image j of $R_{m,n}$. Consider the event set $\{(i \to j|E_j) \mid i \in [m]\}$ together with the probability distribution $Pr[i \to j \mid E_j] = \frac{\binom{j-1}{i-1}\cdot\binom{n-j}{m-i}}{\sum_{i\in[m]} \binom{j-1}{i-1}\cdot\binom{n-j}{m-i}} =$

$\frac{\binom{j-1}{i-1}\cdot\binom{n-j}{m-i}}{\binom{n-1}{m-1}}$. *The information entropy of the probability distribution implies the information kept from the adversary if the adversary views j, which is* $H_j = -\sum_{i\in[m]} Pr[i \to j \mid E_j]\cdot \log Pr[i \to j \mid E_j] = -\sum_{i\in[m]} \frac{\binom{j-1}{i-1}\cdot\binom{n-j}{m-i}}{\binom{n-1}{m-1}} \cdot \log \frac{\binom{j-1}{i-1}\cdot\binom{n-j}{m-i}}{\binom{n-1}{m-1}}$. *Let H be the average remained information if the adversary views a cipher text. Then* $H = -\sum_{j\in[n]} Pr[E_j]\cdot H_j = -n^{-1}\cdot\sum_{j\in[n]}\sum_{i\in[m]} \frac{\binom{j-1}{i-1}\cdot\binom{n-j}{m-i}}{\binom{n-1}{m-1}} \cdot \log \frac{\binom{j-1}{i-1}\cdot\binom{n-j}{m-i}}{\binom{n-1}{m-1}} = HG(m,n)$. *According to Proposition 2.3, $\exists\ 0 < c < 1$ and $m_c > 0$ such that $HG(m,n) \geq c\cdot\log m$ for $n \geq m^2$ and $m \geq m_c$.* $\square$

4.2 General Situation

If there are chosen plain text attacks, the chosen plain texts cut the domain into segments. In order to apply Proposition 4.1 to those segments, we need to prove Lemma 4.2, which analyzes the relation between the distance of plain texts and the distance of the corresponding cipher texts encrypted by $R_{m,n}$. It shows that with high probability, the distance of two cipher texts will be larger than the square distance of the corresponding plain texts, when $n \geq m^3$.

Lemma 4.2 *Suppose that $n \geq m^3$. For all x in $[m]$ and $1 \leq \delta \leq m-x$, $Pr[R_{m,n}(x+\delta) - R_{m,n}(x) < \delta^2 : R_{m,n} \xleftarrow{\$} SIF_{m,n}] \leq \frac{1}{m}$.*

Proof. *Let $R_{m,n}(x) = y$ and $R_{m,n}(x+\delta) = y+\delta'$. There are $\binom{y-1}{x-1}\cdot\binom{\delta'-1}{\delta-1}\cdot\binom{n-y-\delta'}{m-x-\delta}$ many functions in $SIF_{m,n}$ satisfy that condition. Therefore* $Pr[R_{m,n}(x+\delta) - R_{m,n}(x) < \delta^2 : R_{m,n} \xleftarrow{\$} SIF_{m,n}] = \sum_{\delta\leq\delta'<\delta^2}\sum_{x\leq y\leq n-m-\delta'+x+\delta} \frac{\binom{y-1}{x-1}\cdot\binom{\delta'-1}{\delta-1}\cdot\binom{n-y-\delta'}{m-x-\delta}}{\binom{n}{m}} \leq \sum_{\delta\leq\delta'<\delta^2} n\cdot\frac{\binom{\delta'-1}{\delta-1}\cdot\binom{n-\delta'-1}{m-\delta-1}}{\binom{n}{m}} = n\cdot\sum_{\delta\leq\delta'<\delta^2}[\frac{\delta'-1}{\delta-1}\cdot\frac{\delta'-2}{\delta-2}\cdots\frac{\delta'-\delta+1}{1}]\cdot[\frac{(n-\delta'-1)\cdots(n-m-\delta'+\delta+1)}{(m-\delta-1)!}\cdot\frac{m!}{n\cdots(n-m+1)}] \leq n\cdot\sum_{\delta\leq\delta'<\delta^2}[\frac{\delta^2-2}{\delta-1}\cdot\frac{\delta^2-3}{\delta-2}\cdot\delta'\cdots\delta']\cdot[\frac{m\cdots(m-\delta)}{n\cdots(n-\delta)}\cdot\frac{(n-\delta'-1)\cdots(n-m-\delta'+\delta+1)}{(n-\delta-1)\cdots(n-m+1)}] \leq m\cdot\sum_{\delta\leq\delta'<\delta^2}[(\delta+1)\cdot(\delta+3)\cdot\delta'^{\delta-3}]\cdot[\frac{(m-1)\cdots(m-\delta)}{(n-1)\cdots(n-\delta)}\cdot\frac{(n-\delta'-1)\cdots(n-m-\delta'+\delta+1)}{(n-\delta-1)\cdots(n-m+1)}] \leq m\cdot\sum_{\delta\leq\delta'<\delta^2}[m\cdot(m+2)\cdot\delta^{2(\delta-3)}]\cdot[\frac{1}{m^2}\cdots\frac{1}{m^2}\cdot 1\cdots 1] \leq m^2\cdot(m+2)\cdot\sum_{\delta\leq\delta'<\delta^2}(m-1)^{2(\delta-3)}\cdot\frac{1}{m^{2\delta}} < (m+2)\cdot(\delta^2-\delta)\cdot(m-1)^{2(\delta-3)}\cdot\frac{1}{m^{2\delta-2}} < \frac{[(m+2)\cdot(m-1)^2]\cdot(m-1)^{2\delta-6}}{m^{2\delta-2}} < \frac{m^3\cdot(m-1)^{2\delta-6}}{m^{2\delta-2}} \leq \frac{m^3\cdot m^{2\delta-6}}{m^{2\delta-2}} < \frac{1}{m}$.
$\square$

Next we prove Theorem 4.3 for the general situation. Specifically, we prove that if the adversary has the knowledge of at most m^a plain text and cipher text pairs, where $0 < a < 1$, then, the upper bound on the probability for the adversary to recover a plain text x is $[1+(m^{-a}+m_c)\cdot m^{-(1-a)\cdot(1-c)}]\cdot 2^{-[(1-a)\cdot c]\cdot\log m}$ when $n \geq m^3$, where c and m_c are constants and $0 < c < 1$ and $m_c > 0$.

Theorem 4.3 *Let A be an adversary who can collect no more than m^a plain text and cipher text pairs, where $0 < a < 1$. Then, $\exists\ 0 < c < 1$ and $m_c > 0$, s.t.* $Pr[A(\{(x_i, R_{m,n}(x_i)), 1 \leq i \leq i_0\}, R_{m,n}(x)) = x : R_{m,n} \xleftarrow{\$} SIF_{m,n}; x_i \xleftarrow{A} [m], 1 \leq i \leq i_0 < m^a; x \xleftarrow{\$} [m]] \leq [1+(m^{-a}+m_c)\cdot m^{-(1-a)\cdot(1-c)}]\cdot 2^{-[(1-a)\cdot c]\cdot\log m}$ *for $n \geq m^3$.*

Proof. *Without loss of generality, we assume that* $x_0 = 0 < x_1 < \cdots < x_{i_0} < x_{i_0+1} = m+1$. *Let* $Dom_j = [x_{j-1}+1, x_j - 1]$ *and* $\delta_j = x_j - x_{j-1} - 1, 1 \le j \le i_0+1$. *Also without loss of generality, we assume that* $\delta_j \ge m_c$ *for* $1 \le j \le u$ *and* $\delta_j < m_c$ *for* $u+1 \le j \le i_0+1$. *Since* $|\bigcup_{u+1\le j\le i_0+1} Dom_j| \le (i_0 - u + 1)\cdot m_c$, $Pr[x \in \bigcup_{u+1\le j\le i_0+1} Dom_j : x \xleftarrow{\$} [m]] \le (i_0 - u + 1)\cdot \frac{m_c}{m} \le i_0 \cdot \frac{m_c}{m} < \frac{m_c}{m^{1-a}}$.

Consider $x \in Dom_j$, $1 \le j \le u$. *According to Lemma 4.2* $Pr[R_{m,n}(x_j - 1) - R_{m,n}(x_{j-1} + 1) < \delta_j^2 : R_{m,n} \xleftarrow{\$} SIF_{m,n}] \le \frac{1}{m}$. *When* $\delta_j \ge m_c$ *and* $R_{m,n}(x_j - 1) - R_{m,n}(x_{j-1} + 1) \ge \delta_j^2$, $Pr[A(R_{m,n}(x)) = x] \le 2^{-c\cdot\log\delta_j}$ *according to Proposition 4.1.*

Let the event $E =$ "$A(\{(x_i, R_{m,n}(x_i)), 1 \le i \le i_0\}, R_{m,n}(x)) = x : R_{m,n} \xleftarrow{\$} SIF_{m,n}; x_i \xleftarrow{A} [m], 1 \le i \le i_0 < m^a; x \xleftarrow{\$} [m]$" *and the event* $E'_j =$ "$R_{m,n}(x_j - 1) - R_{m,n}(x_{j-1} + 1) \ge \delta_j^2 : R_{m,n} \xleftarrow{\$} SIF_{m,n}$". *Hence* $Pr[E] = \sum_{1\le j\le i_0+1} Pr[E|x \in Dom_j]\cdot Pr[x \in Dom_j] = \sum_{1\le j\le u} Pr[E|x \in Dom_j]\cdot Pr[x \in Dom_j] + \sum_{u+1\le j\le i_0+1} Pr[E|x \in Dom_j]\cdot Pr[x \in Dom_j] \le \sum_{1\le j\le u} Pr[E|x \in Dom_j]\cdot \frac{\delta_j}{m} + \sum_{u+1\le j\le i_0+1} Pr[x \in Dom_j] \le \sum_{1\le j\le u}(Pr[E|E'_j, x \in Dom_j]\cdot Pr[E'_j] + Pr[E|\bar{E}'_j, x \in Dom_j]\cdot Pr[\bar{E}'_j])\cdot \frac{\delta_j}{m} + Pr[x \in \bigcup_{u+1\le j\le i_0+1} Dom_j] \le \sum_{1\le j\le u}(Pr[E|E'_j, x \in Dom_j] + Pr[\bar{E}'_j])\cdot \frac{\delta_j}{m} + \frac{m_c}{m^{1-a}} \le \sum_{1\le j\le u}(2^{-c\cdot\log\delta_j} + \frac{1}{m})\cdot\frac{\delta_j}{m} + \frac{m_c}{m^{1-a}} = \sum_{1\le j\le u}\frac{\delta_j^{1-c}}{m} + \sum_{1\le j\le u}\frac{\delta_j}{m^2} + \frac{m_c}{m^{1-a}}$.

Since $\sum_{1\le j\le u}\delta_j \le m$ *and* $u \le i_0 < m^a$, *we have* $\sum_{1\le j\le u}\frac{\delta_j^{1-c}}{m} \le \sum_{1\le j\le u}\frac{(\frac{m}{u})^{1-c}}{m} = (\frac{u}{m})^c < (\frac{m^a}{m})^c = m^{-(1-a)\cdot c}$. *We also have* $\sum_{1\le j\le u}\frac{\delta_j}{m^2} < \frac{m}{m^2} = \frac{1}{m}$. *Therefore* $Pr[E] \le m^{-(1-a)\cdot c} + \frac{1}{m} + \frac{m_c}{m^{1-a}} = m^{-(1-a)\cdot c} + \frac{(m^{-a}+m_c)}{m^{1-a}} = [1 + (m^{-a} + m_c)\cdot m^{-(1-a)\cdot(1-c)}]\cdot m^{-(1-a)\cdot c} = [1 + (m^{-a} + m_c)\cdot m^{-(1-a)\cdot(1-c)}]\cdot 2^{-[(1-a)\cdot c]\cdot\log m}$ *for* $n \ge m^3$. □

In Theorem 4.3, $[1 + (m^{-a} + m_c)\cdot m^{-(1-a)\cdot(1-c)}]\cdot 2^{-[(1-a)\cdot c]\cdot\log m}$ is a trivial upper bound for $m < m_c$ because $[1+(m^{-a}+m_c)\cdot m^{-(1-a)\cdot(1-c)}]\cdot 2^{-[(1-a)\cdot c]\cdot\log m} > m_c\cdot m^{-(1-a)\cdot(1-c)}\cdot 2^{-[(1-a)\cdot c]\cdot\log m} = m_c\cdot m^{-(1-a)\cdot(1-c)}\cdot m^{-(1-a)\cdot c} = m_c\cdot m^{-(1-a)} > m_c\cdot m_c^{-(1-a)} = m_c^a > 1$. For sufficiently large m, the dominant term of the upper bound is $2^{-[(1-a)\cdot c]\cdot\log m}$ and, hence, $(1-a)\cdot c$ determines the security level of $R_{m,n}$ under chosen plain text attack. The sub-term $1-a$ reflects the influence of chosen plain text attacks and the sub-term c reflects the amount of information leak for a plain text encrypted by $R_{m,n}$ without chosen plain text attacks.

5 Conclusions and Further Research

In this paper we analyze the security of the *OPE* encryption scheme $SE_{m,n}$ and give the upper bound on the probability for the adversary to recover the plain text encrypted by $SE_{m,n}$ under chosen plain text attacks. There are still issues that require further studies in the future, including: (1) It is desirable to compute a tighter upper bound for $SE_{m,n}$ on the probability for the adversary to recover a plain text under chosen plain text attacks; (2) It is also desirable to compute the corresponding lower bound; and (3) The securities of $SE_{m,n}$ for domains with non-uniform data distributions can be analyzed.

References

[1] R. Agrawal, J. Kiernan, R. Stikant, and Y. Xu, *Order-preserving encryption for numeric data*, *ACM SIGMOD International Conference on Management of Data* 2004, pp. 563-574.

[2] Ashwin Swaminathan, Yinian Mao, Guan-Ming Su, Hongmei Gou, Avinash L. Varna, Shan He, Min Wu, Douglas W. Oard, *Confidentiality-preserving rank-ordered search*, *Workshop On Storage Security And Survivability*, Oct. 2007, pp. 7-12.

[3] G. Amanatidis, A. Boldyreva and A. O'Neill, *Provably-secure schemes for basic query support in outsourced databases*, *Working Conference on Data and Applications Security*, 2007, pp. 14-30.

[4] G. Bebek. *Anti-tamper database research: Inference control techniques*, *Technical Report EECS 433 Final Report*, Case Western Reserve University, November 2002.

[5] M. Bellare, A. Boldyreva, and A. O'Neill, *Deterministic and efficiently searchable encryption*, *CRYPTO* 2007, pp. 535-552.

[6] M. Bellare, M. Fischlin, A. O'Neill, and T. Ristenpart, *Deterministic encryption: Definitional equivalences and constructions without random oracles*, *CRYPTO* 2008, pp. 360-378.

[7] A. Boldyreva, S. Fehr, and A. O'Neill, *On notions of security for deterministic encryption, and efficient constructions without random oracles*, *CRYPTO* 2008, pp. 335-359.

[8] A. Boldyreva, N. Chenette, Y. Lee, A. O'Neill, *Order-preserving symmetric encryption*, *Eurocrypt* 2009, pp. 224-241.

[9] D. Boneh and B. Waters, *Conjunctive, subset, and range queries on encrypted data*, *TCC* 2007, pp. 535-554.

[10] Sergei Evdokimov and Oliver G unther, *Encryption techniques for secure database outsourcing*, *ESORICS* 2007, pp. 327-342.

[11] S. C. Gultekin Ozsoyoglu,David Singer, *Anti-tamper databases: Querying encrypted databases*, *Conference on Database and Applications Security*, August 2003.

[12] H. Hacigumus, B. R. Iyer, C. Li, and S. Mehrotra, *Executing SQL over encrypted data in the database-service-provider model*, *ACM SIGMOD Conference on Management of Data*, June 2002.

[13] M. Halloush and M. Sharif, *Global heuristic search on encrypted data (GHSED)*, *International Journal of Computer Science Issues (IJCSI)*, Vol. 1, 2009, pp13-17.

[14] Nancy C. Lee and Jim Finn, *Oracle buys PeopleSoft*, http://www.oracle.com/corporate/press/2004_dec/acqisition.html.

[15] J. Li, E. R. Omiecinski, *Efficiency and security trade-off in supporting range queries on encrypted databases*, *Data and Applications Security* 2005, pp. 69-83.

[16] E. Shi, J. Bethencourt, T-H. H. Chan, D. Song, and A. Perrig, *Multi-dimensional range query over encrypted data*, *Symposium on Security and Privacy* 2007, pp. 350-364.

[17] Tingjian Ge, S. Zdonik, *Fast, secure encryption for indexing in a column-oriented DBMS*, *International Conference on Data Engineering*, 2007, pp. 676-685.

[18] Dawn Xiaoding Song, D. Wagner, A. Perrig, *Practical techniques for searches on encrypted data*, *IEEE Symposium on Security and Privacy*, pp. 44-55, 2000.