A Ciphertext-Size Lower Bound for Order-Preserving Encryption with Limited Leakage

David Cash and Cong Zhang

Department of Computer Science, Rutgers University

Abstract. We consider a recent security definition of Chenette, Lewi, Weis, and Wu for order-revealing encryption (ORE) and order-preserving encryption (OPE) (FSE 2016). Their definition says that the comparison of two ciphertexts should only leak the index of the most significant bit on which the differ. While their work could achieve ORE with short ciphertexts that expand the plaintext by a factor ≈ 1.58, it could only find OPE with longer ciphertxts that expanded the plaintext by a security-parameter factor. We give evidence that this gap between ORE and OPE is inherent, by proving that *any* OPE meeting the information-theoretic version of their security definition (for instance, in the random oracle model) must have ciphertext length close to that of their constructions. We extend our result to identify an abstract security property of any OPE that will result in the same lower bound.

Keywords. Symmetric Encryption, Searchable Encryption, Lower Bound

1 Introduction

To enable fast operations on encrypted databases, several variants of encryption have been suggested that trade security or efficiency for processing functionality on the server. Amongst the suggested constructions, *order-revealing encryption (ORE)* and its special case *order-preserving encryption (OPE)* [1, 3, 4] have seen deployments in products[1] and usage in applied research [12, 14, 13]. ORE schemes, which are the subject of the present paper, are symmetric key encryption schemes $\mathcal{E}$ such that, given ciphertexts $\mathcal{E}_K(x), \mathcal{E}_K(y)$ for messages x, y, one can decide if $x < y$ or not without the decryption key. OPE schemes are the subset of ORE schemes for which the ciphertexts themselves are numbers that can be compared (so $\mathcal{E}_K(x) < \mathcal{E}_K(y) \iff x < y$).

A typical application of ORE is in databases, where one party encrypts numeric columns of a database table. Later, to issue a range query on the column, that party encrypts the endpoints of the range and requests all ciphertexts between them, an operation that can be processed by anyone who holds the

[1] e.g. `https://www.skyhighnetworks.com`, `https://www.ciphercloud.com/`, SAP's SEEED `https://www.sics.se/sites/default/files/pub/andreasschaad.pdf`, `https://www.bluecoat.com/` and Cipherbase [2]

encrypted column. In these settings, OPE is preferable because it can more easily be added to a database application, as the server can be oblivious to the fact that encryption is used at all. With more general ORE schemes, one needs to implement the specialized comparison operation in the database, which can be inconvenient or impossible, for instance when adding encryption to legacy systems.

This work studies the *ciphertext length* of any OPE construction achieving a certain new security notion recently given by a recent work of Chenette et al. [6] (we refer to this work as CLWW below). This notion is currently the best known security property for OPE that can be implemented and deployed. In particular, it results in strictly better security when combined with existing OPE via double-encryption. It seems likely that deployments using OPE (like those mentioned above) will be extended to use CLWW OPE if possible. And although recent attacks have shown that existing OPE is insecure in many contexts [8, 11], it will likely continue to be used in practice in scenarios where the attacks do not apply.

CLWW constructed ORE with their security notion that has ciphertext length $\log_2(3) \cdot m \approx 1.58m$ bits, where m is the plaintext length, and showed how to convert their scheme to the more convenient OPE, but at the cost of increasing the ciphertext length to λm, where λ is the security parameter. This means that achieving OPE comes at a cost of increasing storage of the column by a factor in the range of 80 to 256, compared to the 1.58 expansion of ORE. Achieving smaller OPE ciphertexts with the same security would be highly desirable if possible, as large plaintext data sizes are often the motivating factor for outsourcing data to untrusted server in the first place. (We note that a different, incomparable ORE security notion of [3] can be achieved with $m + 1$ bit ciphertexts, although this fact will not be used in our work below.)

Below we give evidence that the large ciphertext size of the OPE in CLWW is inherent, by proving that any scheme meeting the information-theoretic version of their security notion must have ciphertexts of length

$$\lambda m - m \log m + m \log e,$$

where again m is the message length, logarithms are base 2, and e is the base of the natural logarithm. This bound shows that CLWW has almost optimal ciphertext size, as it has leading term λm instead of $(\lambda - \log m)m$.

In the remainder of this section we describe the prior work on ORE in more detail, and then sketch our results.

ORE SECURITY. It is immediate that an ORE scheme cannot be semantically secure against passive attacks, because one can compute information about plaintexts. But meaningful and formally-defined security targets for ORE have been suggested, starting with the work of [3]. This work defined two notions, the of which was a *ideal* ORE security that requires all plaintext information except order to be hidden. They also showed that no efficient OPE scheme (in particular, one with $\text{poly}(\lambda, m)$-size ciphertexts) could achieve ideal security. However, it was later shown [5] that ideal security for ORE is achievable using crypto-

graphic multilinear pairings [9] or indistinguishability obfuscation [10]. This was apparently the first separation of OPE and ORE as primitives.

Motivated by the lack of a practical ideal construction, Boldyreva et al. [3] investigated a particular weaker notion called ROPF[2]. It was later shown [4] that ROPF-secure ciphers allow a passive adversary to compute the most-significant half of the bits of a random message with high probability, which may be too weak for some applications. The notion was however instantiated with fast blockcipher-based constructions under standard assumptions.

The recent CLWW work [6] introduced a different notion of security for ORE and demonstrated that it is stronger than ROPF-security by certain measures. In particular, that work gave a construction of ORE that could provably hide all but a logarithmic number of bits of a random plaintext. Moreover, the construction is simple to implement and uses only a blockcipher and standard assumptions. The CLWW security notion allows an adversary, given ciphertexts $\mathcal{E}_K(x), \mathcal{E}_K(y)$, to learn the index of the most significant bit on which x and y differ. As mentioned above, the ORE version of their construction has ciphertext size $\approx 1.58m$ while the OPE version has ciphertext size λm.

Our result. For technical reasons discussed below, we consider an information theoretic version of CLWW security, which requires the same security but against unbounded adversaries. The CLWW construction achieves this notion in the random oracle model, and we show that their construction is essentially optimal in terms of ciphertext length. Thus their large overhead in converting their construction from ORE to OPE is inherent, and should OPE with lower storage overhead be required, one will have to investigate other security notions for OPE.

We also generalize our lower bound to apply to any OPE with a new security notion that we call *inner-distance indistinguishability*. While not necessarily interesting as a security goal on its own (one would prefer something stronger), it encapsulates a property that must be avoided in order to build OPE with $O(m)$ size ciphertexts.

Our techniques start from first principles regarding when relations between random variables force their distributions to have large statistical distance. We sketch our proof in Section 4. We note that the *big-jump* attack of Boldyreva et al. [4] proves an exponential lower bound on ideal OPE, and bears some resemblance to our attack. But our attack treats a different and weaker security notion and obtains a fine-grained, polynomial lower bound.

Information-theoretic versus computational security. We attempted to prove our result for any computationally-CLWW-secure ORE scheme, but our techniques do not seem suited to this case. An information-theoretic bound, however, applies to any construction secure in the random-oracle model and includes the CLWW construction. Moreover, if a scheme uses a PRF as its only

[2] We will not need this definition in this paper. Roughly, ROPF security requires that a deterministic cipher be indistinguishable from a *random order preserving function* with the same domain and range.

cryptographic component, then our lower bound applies to a version of that scheme that uses a random-oracle in place of the PRF and thus to the original as well. We are unaware of any technique for building computationally-CLWW-secure OPE that circumvents our bound, and we conjecture that a ciphertext length lower bound also holds in the computational case.

ORGANIZATION. In Section 2 we recall definitions for ORE/OPE syntax and security and in Section 3 we recall the specific security notion that we study. In Section 4 we state our lower bound and sketch its proof, which is given in Sections 5, 6, 7. Finally in Section 8 we show to generalize our result to an abstract security property.

2 Preliminaries

NOTATION AND BASIC RESULTS. We always use λ to denote the security parameter. For non-negative integers $a \leq b$ we write $[a, b]$ for the set $\{a, a+1, \ldots, b\}$, $[n]$ for the set $\{1, \ldots, n\}$, and $[n]'$ for the set $\{0, 1, \ldots, n\}$. If X_1, X_2 are r.v.s, we let

$$\Delta(X_1, X_2) = \frac{1}{2} \sum_k |\Pr[X_1 = k] - \Pr[X_2 = k]|$$

denote their statistical distance. We will use the following well-known *data processing lemma* (c.f. [7]) in our proof.

Lemma 1. *Let X and Y be r.v.s, and f be any function that includes the support of X and Y in its domain. Then $\Delta(f(X), f(Y)) \leq \Delta(X, Y)$.*

For a randomized algorithm $\mathcal{A}$ we write $y \xleftarrow{\$} \mathcal{A}(w)$ to denote running $\mathcal{A}$ on input w, and letting y be the random variable denoting its output. If $\mathcal{A}$ is deterministic, we denote $y \leftarrow \mathcal{A}(w)$ to denote running $\mathcal{A}$ and letting y be its output.

We write $\mathbf{1}(x < y)$ to mean 1 if $x < y$ and 0 otherwise.

ORE AND OPE. An *ORE scheme* Π is a tuple of algorithms $(\mathcal{K}, \mathcal{E}, \mathcal{C})$ for key generation, encryption, and comparison respectively, and always has an associated message space $\{0,1\}^m$ and ciphertext space $\{0,1\}^n$. The key generation algorithm $\mathcal{K}$ is randomized, and on input 1^λ, outputs a key K. The encryption algorithm $\mathcal{E}$ is deterministic and takes as input a key K and message $x \in \{0,1\}^m$ and outputs a ciphertext $c \leftarrow \mathcal{E}_K(x)$. The comparison algorithm takes as input two ciphertexts c_1, c_2 generated with the same K on messages x_1, x_2 and outputs a bit b.

We assume that all ORE schemes in this paper are *correct*, meaning that for all λ, keys K in the support of $\mathcal{K}(1^\lambda)$, and all $x, y \in \{0,1\}^m$, $\mathcal{C}(\mathcal{E}_K(x), \mathcal{E}_K(y))$ outputs $\mathbf{1}(x < y)$. Note that this allows testing if $x = y$ by running the comparison algorithm twice.

When an scheme ORE scheme Π has a canonical comparison algorithm $\mathcal{C}$ that directly compares its inputs as numbers in $[2^n - 1]'$, we say that the scheme is an *order-preserving encryption (OPE)* scheme. In this case we omit the comparison algorithm and write $\Pi = (\mathcal{K}, \mathcal{E})$.

ORE SECURITY. Chenette et al. [6] gave a simulated-based definition for ORE security that used a leakage profile $\mathcal{L}$ as a parameter, where $\mathcal{L}$ is an efficient algorithm. We will use a weaker non-interactive indistinguishability-based version of their definition for our lower bounds (which makes our result stronger).

For an ORE scheme $\Pi = (\mathcal{K}, \mathcal{E}, \mathcal{C})$, leakage profile $\mathcal{L}$, and adversary $\mathcal{A}$ we consider the following game:

Game $\text{ORE}_{\Pi,\mathcal{L},\mathcal{A}}(\lambda)$:

$K \xleftarrow{\$} \mathcal{K}(1^\lambda)$
$(\mathbf{m}_0, \mathbf{m}_1, s) \xleftarrow{\$} \mathcal{A}(\lambda)$
If $\mathcal{L}(\mathbf{m}_0) \neq \mathcal{L}(\mathbf{m}_1)$ then output 0.
For $i = 1, \ldots q$: $\mathbf{c}[i] \leftarrow \mathcal{E}_K(\mathbf{m}_b[i])$
$b' \xleftarrow{\$} \mathcal{A}(s, \mathbf{c})$
If $b' = b$ then output 1, Else output 0,

We define the *$\mathcal{L}$-advantage of $\mathcal{A}$ against Π* to be

$$\mathbf{Adv}^{\text{ore}}_{\Pi,\mathcal{L},\mathcal{A}}(\lambda) = 2\Pr[\text{ORE}_{\Pi,\mathcal{L},\mathcal{A}}(\lambda) = 1] - 1.$$

We say that Π is *$\mathcal{L}$-computationally secure* if for all efficient $\mathcal{A}$, $\mathbf{Adv}^{\text{ore}}_{\Pi,\mathcal{L},\mathcal{A}}(\lambda)$ is a negligible function i.e. is $o(1/\operatorname{poly}(\lambda))$. We say that Π is *$\mathcal{L}$-statistically-secure* if the same condition holds for all (unbounded, wlog deterministic) adversaries $\mathcal{A}$.

We recall, as an example, that the *ideal* leakage profile only leaks order. Formally, this is

$$\mathcal{L}_{\text{ideal}}(m_1, \ldots, m_q) = \{(i, j, \mathbf{1}(m_i < m_j)) \ : \ 1 \leq i < j \leq q\}.$$

3 CLWW Security and Constructions

In this section we recall and discuss the CLWW leakage profile and constructions.

CLWW LEAKAGE. CLWW considered the following leakage profile $\mathcal{L}_{\text{clww}}$. On input $\mathbf{x} = (x_1, \ldots, x_q) \in (\{0,1\}^m)^q$, the leakage profile is defined by

$$\mathcal{L}_{\text{clww}}(x_1, \ldots, x_q) := \{(i, j, \mathsf{ind}_{\mathsf{diff}}(x_i, x_j), \mathbf{1}(x_i < x_j)) \ : \ 1 \leq i < j \leq q\},$$

where $\mathsf{ind}_{\mathsf{diff}}(x_i, x_j) \in \{1, \ldots, m+1\}$ is the left-most bit on which x_i and x_j differ, or $m+1$ if they are equal. Compared to the ideal profile, only the $\mathsf{ind}_{\mathsf{diff}}(x_i, x_j)$ indices are extra leakage.

The intuition for the leakage is that, when comparing two numbers, and adversary will learn the length of the longest common prefix, and also which is larger. This information combines to reveal one bit of each of the plaintexts.

THE CLWW ORE AND OPE CONSTRUCTIONS. Our results will not need the CLWW construction, but it provides intuition for the lower bound and we recall it now, starting with a basic ORE construction $\Pi_{\text{clww-ore}}$ and then describing

an ORE variant with shorter ciphertexts, and how to build OPE $\Pi_{\text{clww-ope}}$. We recall a version that is slightly different from theirs in that it is *perfectly* correct.

The scheme $\Pi_{\text{clww-ore}} = (\mathcal{K}^{\text{ore}}, \mathcal{E}^{\text{ore}}, \mathcal{C}^{\text{ore}})$ uses a PRF

$$F : \{0,1\}^\lambda \times ([m] \times \{0,1\}^m) \to (\{0,1\}^\lambda \setminus \{1^\lambda\}).$$

Thus the input domain of F is $[m] \times \{0,1\}^m$, and it outputs a λ-bit string that is assumed to never be 1^λ (of course we can modify any PRF so that this is true without affecting asymptotic security).

- Key generation $\mathcal{K}^{\text{ore}}(1^\lambda)$ outputs a random PRF key $K \xleftarrow{\$} \{0,1\}^\lambda$.
- Encryption $\mathcal{E}^{\text{ore}}_K(x)$, on input a message $x \in \{0,1\}^m$, the computes for each $i = 1, \ldots, m$ the value

$$u_i = F(K, i \,\|\, x[1, \ldots, i-1] \,\|\, 0^{m-i+1}) + x[i], \tag{1}$$

  where the addition is done by interpreting the bitstrings as members of $\{0, \ldots, 2^\lambda - 1\}$. Encryption outputs $(u_1, \ldots, u_m)$.
- The comparison algorithm $\mathcal{C}^{\text{ore}}((u_1, \ldots, u_m), (u'_1, \ldots, u'_m))$ takes as input two ciphertexts. It finds the smallest i such that $u_i \neq u'_i$, and it outputs 1 if $\mathbf{1}(u_i < u'_i)$.

Correctness follows by observing that the u_i will be equal until the u_i, u'_i corresponding to the first differing bit in the plaintexts. At that position, u_i and u'_i will differ by 1 (additively) and the smaller plaintext has the smaller value. CLWW proved that $\Pi_{\text{clww-ore}}$ (and the variants below) are $\mathcal{L}_{\text{clww}}$-secure, assuming that F is a PRF. It is straightforward to derive from their proof that $\Pi_{\text{clww-ore}}$ is also statistically-secure with the same leakage profile in the random-oracle model.

Conversion to OPE. Chenette et al. showed how to convert this construction to an OPE scheme $\Pi_{\text{clww-ope}}$ by simply concatenating the members of a ciphertext to form a bitstring in $\{0,1\}^{\lambda m}$ that is interpreted as a number for comparison. This scheme is perfectly correct because of our assumption that F never outputs the all-ones string, and thus the addition in (1) will never wrap modulo 2^λ.

Compressing ORE ciphertexts. Chenette et al. showed that one can modify $\Pi_{\text{clww-ore}}$ to have shorter ciphertexts by changing instead using a PRF F' with range only $\{0,1,2\}$, so

$$F' : \{0,1\}^\lambda \times ([m] \times \{0,1\}^m) \to \{0,1,2\}.$$

Now encryption uses F', and for $i = 1, \ldots, m$ computes

$$u_i = F(K, i \,\|\, x[1, \ldots, i-1] \,\|\, 0^{m-i+1}) + x[i] \mod 3. \tag{2}$$

It outputs the vector $(u_1, \ldots, u_n) \in \{0,1,2\}^m$.

Comparison now takes as input $(u_1, \ldots, u_m)$ $(u'_1, \ldots, u'_m)$. As before, it finds the first i such that $u_i \neq u'_i$. But now it outputs 1 if $u'_i = u_i + 1 \mod 3$, and otherwise it outputs 0.

A ciphertext for an m-bit input is now a vector in $\{0,1,2\}^m$, which can be represented using $log_2(3)m + O(1) \approx 1.58m$ bits.

4 Lower Bound Statement and Proof Sketch

We can now state our lower bound formally.

Theorem 2. *Suppose $\Pi = (\mathcal{K}, \mathcal{E}, \mathcal{C})$ is an order-preserving encryption scheme with associated message space $\{0,1\}^m$ and ciphertext space $\{0,1\}^n$, and that Π is $2^{-\lambda}$-$\mathcal{L}_{\text{clww}}$-statistically-secure. Then we have*

$$n \geq \lambda m - m \log m + m \log e$$

In any practical OPE scenario we are aware of, we have $\log m - \log e < \lambda$ and thus our bound is nontrivial. For example, considering the message space is 40 bytes, $\log m - \log e = \log 320/e < 7$, while in real world encryption, the secure parameter is always set to be 80 or larger.

NOTATION FOR THE PROOF. To explain why this theorem is true we start with a change of notation that is more convenient for the underlying statistical problem. We will freely treat a string $i \in \{0,1\}^m$ as a member of $[2^m-1]' = \{0, \ldots, 2^m-1\}$ when convenient (and similarly for strings in $\{0,1\}^n$). For each $i \in \{0,1\}^m$ we define a random variable X_i by $X_i = \mathcal{E}_K(i)$, where $K \xleftarrow{\$} \mathcal{K}(1^\lambda)$. These random variables are dependent, and perfect correctness implies that $X_0 < X_1 < \cdots < X_{2^m-1}$ with probability one (here we are treating the X_i as numbers).

Now we consider what the ϵ-$\mathcal{L}_{\text{clww}}$-statistical security implies about our r.v.s $X_0, \ldots, X_{2^m-1}$. For every possible pair of vectors of messages $\mathbf{m}_0, \mathbf{m}_1$ that does not automatically lose the game because of the leakage requirement, we get a condition about the statistical distance of the distributions of two tuples of random variables. For instance, if the adversary requests singleton vectors $\mathbf{m}_0 = i$ or $\mathbf{m}_1 = j \in \{0,1\}^m$ then the leakage $\mathcal{L}_{\text{clww}}(i) = \mathcal{L}_{\text{clww}}(j) = \emptyset$, so we must have that

$$\Delta(X_i, X_j) \leq \varepsilon$$

for every i, j. More generally, for any two vectors $\mathbf{i} = (i_1, \ldots, i_q)$ and $\mathbf{j} = (j_1, \ldots, j_q)$ in $(\{0,1\}^m)^q$ with $\mathcal{L}_{\text{clww}}(\mathbf{i}) = \mathcal{L}_{\text{clww}}(\mathbf{j})$, we must have

$$\Delta((X_{i_1}, \ldots, X_{i_q}), (X_{j_1}, \ldots, X_{j_q})) \leq \epsilon.$$

Thus we need to understand which $\mathbf{i}, \mathbf{j}$ satisfy $\mathcal{L}_{\text{clww}}(\mathbf{i}) = \mathcal{L}_{\text{clww}}(\mathbf{j})$. Fortunately, our proof will only require inputs of a particular structure. We observe that the following qualify for $t = 0, \ldots, m-1$:

$$\mathbf{i} = (0, 2^{t+1} - 1) \quad \text{and} \quad \mathbf{j} = (2^t - 1, 2^t).$$

In binary, $\mathbf{i}$ is $(0^m, 0^{m-t-1}1^{t+1})$ and $\mathbf{j}$ is $(0^{m-t}1^t, 0^{m-t-1}10^t)$. In both cases, the most significant differing bit is in the $t+1$-st least significant position (and the messages are in the same order), so the leakage in the same.

But why should this choice be useful? It represents the most extreme cases of two "distant" plaintexts and two "close" plaintexts that must appear indistinguishable. At a very high level, the scheme must "waste" a lot of its ciphertext

space in order to make pairs like this appear indistinguishable. This is because the **i** side must have ciphertexts that are far apart (by roughly 2^{t+1}) simply because correctness forces many ciphertexts to be between X_0 and $X_{2^{t+1}-1}$, namely $X_1, X_2, \ldots, X_{2t+1-2}$. In order to appear indistinguishable, X_{2^t-1} and X_{2^t} must also be far apart, with no other ciphertexts between them (again by correctness). Moreover, as t grows we get a *nested* sequence of pairs, where the space wasted by the previous pair force the next to waste even more.

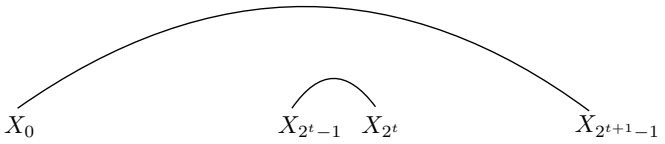

Fig. 1: Two pairs of r.v.s that are required to be indistinguishable by the security definition. The top arc represents the gap G_1 and the bottom arc represents the gap G_2.

Our proof will argue that this wasted space grows to the quoted bound. We consider the nested sequence of these tuples above, and then proceed by induction to show that a large ciphertext-space is needed for security. The key step in our induction is that, since the tuples $(X_0, X_{2^{t+1}-1})$ and (X_{2^t-1}, X_{2^t}) must have statistical distance at most ϵ, then their *gaps*

$$G_1 = X_{2^{t+1}-1} - X_0 \quad \text{and} \quad G_2 = X_{2^t} - X_{2^t-1}$$

must also satisfy $\Delta(G_1, G_2) \leq \varepsilon$ by the data processing inequality. But the gap measured by G_2 is a subset of the gap measured by G_1, so $G_2 < G_1$. In fact, as we show via induction on t, G_2 must often be much less than G_1 (since G_1 contains the gap from X_{2^t-1} and X_0, which is the previous step of the induction). Using this fact, we apply the following lemma that is proved in Section 6.

Lemma 3. *For any two variables $X \geq Y \in [N-1]'$, and distinct positive integers $d_1, \ldots, d_k$ such that $Pr[X = Y + d_i] = p_i$, we have*

$$\Delta(X, Y) \geq \frac{\sum_{i=1}^{k} p_i \cdot d_i}{N-1}.$$

Intuitively, this lemma says that if one of the random variables is often much bigger than the other, then they must have large statistical distance.

Contrast with big jump. The *big jump* attack of [4] gave a ciphertext-size lower bound for any ideal OPE. With ideal ORE, *every* pair of two random variables $X_{i_1} < X_{i_2}$ and $X_{j_1} < X_{j_2}$ must be indistinguishable, which gives the attack more flexibility and results in an exponential bound (without resorting to recursion). Instead our bound works with a particular nested set of m pairs, with each step using a pair to increase the bound by roughly λ bits.

5 Proof of Theorem 2

We start with an additional technical lemma (proved in Section 7), and then give the proof.

Lemma 4. *Let $X > Y \in [N-1]'$ be random variables such that $\Delta(X,Y) \leq \delta$. Let $i \geq 1$ and assume if for all $q \in [0,1]$, $\Pr[X > Y + \frac{(1-q)^i}{\delta^i \cdot i!}] \geq q$. Then for all $q \in [0,1]$ we have*

$$\Pr[X > \frac{(1-q)^{i+1}}{\delta^{i+1}(i+1)!}] \geq q.$$

This lemma says that if X is often much larger Y, but also has small statistical distance, then the support of X must include some very large elements, and in fact X concentrates a significant portion of its mass on those large elements. The proof of this lemma (and the proof of the theorem) depends on Lemma 3 from above. We remark that it is crucial that we have the same probability q in the lemma assumption and conclusion, and achieving this requires a delicate argument. A weaker conclusion, where q changes, is more easily achieved using a Markov-type argument (and indeed earlier versions of this paper did exactly this, resulting in a weaker bound).

5.1 Proof

Let $\Pi = (\mathcal{K}, \mathcal{E})$ be an OPE scheme with associated message space $\{0,1\}^m$ and ciphertext space $\{0,1\}^n$, and assume Π is $2^{-\lambda}$-$\mathcal{L}_{\text{clww}}$-statistically-secure.

Below, for $i \in [2^m - 1]'$, we let $X_i = \mathcal{E}_K(i)$ where $K \xleftarrow{\$} \mathcal{K}(1^\lambda)$ as in the proof sketch. That is, the X_i are dependent random variables that represent the encryption of message i under a random key. Note that $X_1 < X_2 < \cdots < X_{2^m-1}$.

We will prove the theorem using following claim. Here, we let $\varepsilon = 2^{-\lambda}$.

Lemma 5. *For $i \in [2^m - 1]'$, let X_i be defined as above. Then for $1 \leq j \leq m$ and $q \in [0,1]$,*

$$\Pr[X_{2^j-1} - X_0 \geq \frac{(1-q)^{j-1}}{\varepsilon^{j-1} \cdot (j-1)!}] \geq q$$

Proof (of Lemma 5). The proof is by induction on j.

Case $j = 1$. This case reduces to $\Pr[X_1 - X_0 \geq 1] = 1$, which is true by the correctness of the scheme.

Case $j \implies j+1$. We need to show that for any $q \in [0,1]$

$$\Pr[X_{2^{j+1}-1} - X_0 \geq \frac{(1-q)^j}{\varepsilon^j \cdot (j)!}] \geq q.$$

By the correctness of the scheme, we have that

$$X_{2^{j+1}-1} - X_0 \geq (X_{2^j} - X_{2^j-1}) + (X_{2^j-1} - X_0) + 1 \tag{3}$$

Now define "gap" random variables $G_1 = X_{2^{j+1}-1} - X_0$ and $G_2 = (X_{2^j} - X_{2^j-1})$. By induction we know that for any $q \in [0,1]$

$$\Pr[X_{2^j-1} - X_0 \geq \frac{(1-q)^{j-1}}{\varepsilon^{j-1} \cdot (j-1)!}] \geq q.$$

Plugging this, and the definitions of G_1, G_2 into (3), we have

$$\Pr[G_1 > G_2 + \frac{(1-q)^{j-1}}{\varepsilon^{j-1} \cdot (j-1)!}] \geq q.$$

Moreover, we know by the ε-$\mathcal{L}_{\text{clww}}$-statistical security of Π and Lemma 1 that $\Delta(G_1, G_2) \leq \varepsilon$.

We now want to apply Lemma 4 to G_1 and G_2, to show that G_1 must be large and then conclude the induction. In the lemma, we set $G_1 = X, G_2 = Y, i = j$, and $\delta = \varepsilon$. The lemma gives

$$\Pr[G_1 > \frac{(1-q)^{j}}{\varepsilon^{j} \cdot (j)!}] \geq q,$$

obtaining the induction step.

We can now complete the proof of Theorem 2. The above lemma with $j = m$ tells us that for any $q \in [0,1]$

$$\Pr[X_{2^m-1} > X_0 + \frac{(1-q)^{m-1}}{\varepsilon^{m-1} \cdot (m-1)!}] \geq q,$$

and thus for any $j \leq D = 1/\varepsilon^{m-1}(m-1)!$,

$$\Pr[X_{2^m-1} > X_0 + j] \geq 1 - ((m-1)! \cdot j)^{1/m-1}\varepsilon$$

and

$$\sum_{\ell=1}^{j} \Pr[X_{2^m-1} = X_0 + \ell] \leq ((m-1)! \cdot j)^{1/m-1}\varepsilon.$$

Besides, we claim $D \leq N-1$, if not, then there exists $q > 0$ such that

$$N - 1 = \frac{(1-q)^{m-1}}{\varepsilon^{m-1} \cdot (m-1)!}$$

referring to

$$\Pr[X_{2^m-1} > X_0 + N - 1] \geq q > 0$$

which contradicts $X_i \in [N-1]'$.

Now we denote $p_\ell = \Pr[X_{2^m-1} = X_0 + \ell]$, and according to Lemma 3, we get that

$$\varepsilon \geq \Delta(X_{2^m-1}, X_0) \geq \frac{\sum_{\ell=1}^{N-1} p_\ell \cdot \ell}{N-1} \tag{4}$$

and

$$\begin{aligned}
\sum_{\ell=1}^{N-1} p_\ell \cdot \ell &= (p_1 + \cdots + p_{N-1}) + (p_2 + \cdots + p_{N-1}) + \cdots + p_{N-1} \\
&\geq 1 + (1 - p_1) + (1 - p_1 - p_2) + \cdots + (1 - p_1 - \cdots - p_{D-1}) \\
&\geq 1 + \sum_{\ell=1}^{D-1} (1 - ((m-1)!\ell)^{\frac{1}{m-1}} \cdot \varepsilon) \\
&= D - (m-1)!^{\frac{1}{m-1}} \cdot \varepsilon \sum_{\ell=1}^{D-1} \ell^{\frac{1}{m-1}} \\
&\geq D - (m-1)!^{\frac{1}{m-1}} \cdot \varepsilon \cdot \int_0^D x^{\frac{1}{m-1}} dx \\
&= \frac{1}{\varepsilon^{m-1}(m-1)!} \cdot \frac{1}{m} = \frac{1}{\varepsilon^{m-1} m!}.
\end{aligned}$$

Returning to (4), we have

$$N - 1 \geq 1/\varepsilon^m m!.$$

By setting $\varepsilon = 2^{-\lambda}$, we get

$$n \geq \lambda m - \log(m!) \geq \lambda m - \log((m/e)^m) = \lambda m - m \log m + m \log e.$$

□

6 Proof of Lemma 3

We recall the lemma.

Lemma 3. *For any two variables $X \geq Y \in [N-1]'$, and distinct positive integers $d_1, \ldots, d_k$ such that $Pr[X = Y + d_i] = p_i$, we have*

$$\Delta(X, Y) \geq \frac{\sum_{i=1}^k p_i \cdot d_i}{N - 1}.$$

The following proof was contributed by colleagues whose names are redacted for anonymity. An earlier version of this work gave a much more complicated proof.

Proof. We will show that one of the distinguishers $\mathcal{D}_i$, $i \in [N-1]$, has the needed advantage, where $\mathcal{D}_i$ is defined as follows: Given input $T \in [N-1]'$, $\mathcal{D}_i$ outputs 1 if an only if $T \geq i$.

The advantage of $\mathcal{D}_i$ is $\delta_i = \Pr[X \geq i] - \Pr[Y \geq i]$. We have that

$$\begin{aligned}\sum_{i=1}^{N-1} \delta_i &= \sum_{i=1}^{N-1} \Pr[X \geq i] - \sum_{i=1}^{N-1} \Pr[Y \geq i] \\ &= \sum_{i=0}^{N-1} \Pr[X \geq i] - \sum_{i=0}^{N-1} \Pr[Y \geq i] = E(X-Y) \geq \sum_{i=1}^{k} p_i d_i.\end{aligned}$$

Thus some δ_i must be at least this sum divided by $N-1$. □

7 Proof of Lemma 4

We first recall the lemma.

Lemma 4. *Let $X > Y \in [N-1]'$ be random variables such that $\Delta(X,Y) \leq \delta$. Let $i \geq 1$ and assume if for all $q \in [0,1]$, $\Pr[X > Y + \frac{(1-q)^i}{\delta^i \cdot i!}] \geq q$. Then for all $q \in [0,1]$ we have*

$$\Pr[X > \frac{(1-q)^{i+1}}{\delta^{i+1}(i+1)!}] \geq q.$$

Proof. Suppose for contradiction that there exists $q^* \in [0,1]$ such that

$$\hat{q} := \Pr[X > t] < q^*,$$

where $t = (1-q^*)^{i+1}/\delta^{i+1}(i+1)!$.

We will show that $\Delta(X,Y) > \delta$, violating the assumption in the lemma. We will prove this by showing the following "truncated" r.v.s W, Z satisfy $\Delta(X,Y) \geq \Delta(W,Z) > \delta$, where W, Z are defined via the joint distribution

$$\Pr[W = a, Z = b] = \begin{cases} \Pr[X = a, Y = b] & \text{if } (a,b) \in [t]^2 \setminus (0,0), \\ \hat{q} & \text{if } (a,b) = (0,0) \\ 0 & \text{otherwise} \end{cases}.$$

According to the definition of (W,Z), we show $\Delta(X,Y) \geq \Delta(W,Z)$. For simplify, we denote

$$p_{i,j} = \Pr[X = i, Y = j];\ p_j = \sum_{k=0}^{t} p_{k,j};\ p_j^* = \sum_{k=t+1}^{N-1} p_{k,j};\ \forall i,j \in [t]$$

and it's obvious to note that for $j \in [t]$: 1) $\Pr[X = j] = \Pr[W = j]$; 2) $\Pr[Z = j] = p_j$; 3) $\Pr[Y = j] = p_j + p_j^*$; 4)$\sum_{k=0}^{t} p_j^* = \sum_{k=t+1}^{N-1}(\Pr[X = k] - \Pr[Y = k])$.

Hence:

$$\begin{aligned}
2\Delta(X,Y) &= \sum_{j=0}^{N-1} |\Pr[X=j] - \Pr[Y=j]| \\
&= \sum_{j=0}^{t} |\Pr[X=j] - \Pr[Y=j]| + \sum_{j=t+1}^{N-1} |\Pr[X=j] - \Pr[Y=j]| \\
&\geq \sum_{j=0}^{t} |\Pr[X=j] - \Pr[Y=j]| + \sum_{j=t+1}^{N-1} (\Pr[X=j] - \Pr[Y=j]) \\
&= \sum_{j=0}^{t} |\Pr[X=j] - \Pr[Y=j]| + \sum_{j=0}^{t} p_j^* \\
&= \sum_{j=0}^{t} |\Pr[W=j] - \Pr[Z=j] - p_j^*| + \sum_{j=0}^{t} p_j^* \\
&\geq \sum_{j=0}^{t} |\Pr[W=j] - \Pr[Z=j]| = 2\Delta(W,Z)
\end{aligned}$$

In the following, it suffices to show that $\Delta(W,Z) > \delta$. We denote $d_j = \Pr[W = Z + j]$. Applying Lemma 3,

$$\Delta(W,Z) \geq \frac{\sum_{\ell=1}^{t} d_\ell \cdot \ell}{t}.$$

We now show that $\sum_{\ell=1}^{t} d_\ell \cdot \ell > \delta t$, completing the proof. Below we use the following technical claim, which we establish below:

Claim. In the notation of the proof, we have the following:

1. $\sum_{\ell=1}^{t} d_\ell = 1 - \hat{q}$,
2. For each j, $\sum_{\ell=1}^{j} d_\ell \leq (i! \cdot j)^{1/i}\delta$,
3. $t \geq \hat{t}$, where $\hat{t} = (1-\hat{q})^i/\delta^i i!$.

Using the claim, we have

$$\begin{aligned}
\sum_{\ell=1}^{t} d_\ell \cdot \ell &\geq \sum_{\ell=1}^{\hat{t}} d_\ell \cdot \ell = (d_1 + \ldots + d_{\hat{t}}) + (d_2 + \ldots + d_{\hat{t}}) + \ldots + (d_{\hat{t}}) \\
&\geq (1-\hat{q}) + ((1-\hat{q}) - d_1) + ((1-\hat{q}) - d_1 - d_2) + \ldots + ((1-\hat{q}) - d_1 - \ldots - d_{\hat{t}-1}) \\
&\geq (1-\hat{q})\hat{t} - \sum_{\ell=1}^{\hat{t}-1} (\ell i!)^{1/i}\delta \\
&\geq (1-\hat{q})\hat{t} - (i!)^{1/i}\delta \int_0^{\hat{t}} x^{1/i} dx \\
&= (1-\hat{q})^{i+1}\hat{t} - (i!)^{1/i}\delta \cdot \frac{i}{i+1}\hat{t}^{\frac{i+1}{i}} = \frac{(1-\hat{q})^{i+1}}{\delta^i (i+1)!} > \delta t.
\end{aligned}$$

We now prove the claim. The first part follows easily from the definition of W, Z. For the second part, we have

$$\sum_{\ell=1}^{j} d_j \leq \sum_{\ell=1}^{j} \Pr[X = Y + \ell] = 1 - \Pr[X > Y + j] \leq (i!j)^{1/i}\delta,$$

where the last inequality follows since $\Pr[X > Y + (1-q)^i/\delta^i i!] \geq q$ holds for all $q \in [0,1]$, and particular $q = (i!j)^{1/i}\delta$.

For the third part of the claim, suppose for contradiction that $t < \hat{t}$. Then

$$\Pr[X > t] \geq \Pr[X > Y + t] \geq 1 - (i!t)^{1/i}\delta > 1 - (i!\hat{t})^{1/i}\delta = \hat{q}.$$

(The second inequality is another application of the condition in the lemma, similar to the proof of the second part.) But this contradicts the definition $\hat{q} = \Pr[X > t]$ and proves the third part of the claim. □

8 Extensions of the Lower Bound

Our lower bound applies to the specific definition achieved by Chenette et al., and it is possible to circumvent the bound by targeting a different, but hopefully satisfactory, notion of security. In this section we identify an abstract property, which we term *inner-distance-indistinguishablity*, for which a similar lower bound applies. Thus, to avoid the bound for OPE with another definition, one must avoid this property, and the authors are not aware of an approach for doing so.

We also show how to apply our proof technique to give an essentially-tight lower bound on the ciphertext length of the "base-d" OPE variants suggested by Chenette et al., which achieve a weakened version of security with shorter ciphertexts.

Inner-distance-indistinguishablity. The following property seems mostly useful as a tool for understanding and generalizing the lower bound, and not as a stand-alone target for OPE security in practice.

Definition 6. *Let $\Pi = (\mathcal{K}, \mathcal{E}, \mathcal{C})$ be an OPE scheme with associated message space M, $d \geq 1$ be an integer, and $\varepsilon > 0$. We say that Π is (statistically)* ε-inner-distance-indistinguishable for width d (denoted ε-IDI_d) *if for all $i < j \in M$ such that $j - i > d$, there exist $k, \ell \in M$ such that*

1. $i \leq k < \ell \leq j$
2. $\ell - k \leq d$
3. $\Delta(D_1, D_2) \leq \varepsilon$, *where $D_1 = \mathcal{E}_K(j) - \mathcal{E}_K(i)$ and $D_2 = \mathcal{E}_K(k) - \mathcal{E}_K(\ell)$ and K is random key.*

Intuitively, ε-IDI_d says that the distance between every encrypted pair of messages must be indistinguishable from the gap of between two encrypted messages which both lie between them, and moreover the latter gap is required to be small, namely d or less.

The CLWW notion implies ε-IDI_1 security. That is, for every pair $i < j$, $\mathcal{E}_K(j) - \mathcal{E}_K(i)$ is distinguishable from $\mathcal{E}_K(k+1) - \mathcal{E}_K(k)$ for some k between i and j (when $d = 1$, we must have $\ell = k + 1$ in the definition).

To see this, fix some i, j, with $j > i+1$, and consider their binary expansions. We may write i in the form $p \,\|\, 0 \,\|\, x$ and j in the form $p \,\|\, 1 \,\|\, y$, where p is the longest common prefix and i and j, and $x, y \in \{0,1\}^L$ for some $L \geq 1$. Then consider

$$k = p \,\|\, 0 \,\|\, 1^L \quad \text{and} \quad \ell = p \,\|\, 1 \,\|\, 0^L.$$

We have that $\ell = k+1$ (treating ℓ, k as numbers), and that either $k \neq i$ or $\ell \neq j$. Moreover the CLWW security notion ensures that the condition of IDI_1 security holds for this choice of k, ℓ.

The following theorem generalizes Theorem 2.

Theorem 7. *Suppose $\Pi = (\mathcal{K}, \mathcal{E}, \mathcal{C})$ is an order-preserving encryption scheme with security parameter λ and associated message space $\{0,1\}^m$ and ciphertext space $\{0,1\}^n$, and Π is $2^{-\lambda}$-IDI_d secure for some $d \geq 1$. Let $m' = m - \lceil \log d \rceil$. Then we have*

$$n \geq \lambda m' - m' \log m' + m' \log e$$

Proof. Let $\Pi = (\mathcal{K}, \mathcal{E}, \mathcal{C})$ be an OPE scheme with the syntax and conditions in the theorem. Below, for $i \in \{0,1\}^m$, we write $X_i = \mathcal{E}_K(i)$, and let m' be as defined in the theorem.

We will show how to carry out the same strategy used in the proof of Theorem 2. We will prove a version of Lemma 5 for a different nested sequence of pairs of messages $(i_j^L, i_j^R)_{j=1}^{m'}$ that we define inductively from m' down to 1 now.

- Base: $i_{m'}^L = 0, i_{m'}^R = 2^m - 1$.
- Step: Given (i_j^L, i_j^R), let $k < \ell$ be the pair between i_j^L and i_j^R guaranteed by IDI_d security. We distinguish two cases:
  1. If $k - i_j^L > i_j^R - \ell$ then set (i_{j-1}^L, i_{j-1}^R) to be (i_j^L, k).
  2. Otherwise, set (i_{j-1}^L, i_{j-1}^R) to (ℓ, i_j^R).

Intuitively, we use the IDI_d security property to find a nested sequence by moving to the "larger" gap at each step, and this continues for at least m' steps. Using this sequence, the rest of the proof of Lemma 5 can be carried out. Finally, the rest of the proof of Theorem 2 can be applied exactly as before. □

Extension to OPE variants. We can also extend our proof of Theorem 2 to the "d-ary" variants of Chenette et al. That construction saved a modest amount of space over the main CLWW construction via additional leakage, which is described via the following leakage profile $\mathcal{L}^d_{\text{clww}}$:

$$\mathcal{L}^d_{\text{clww}}(x_1, \ldots, x_q) := \{(i, j, \mathsf{ind}^{(d)}_{\mathsf{diff}}(x_i, x_j), \mathbf{1}(x_i < x_j)) \ : \ 1 \leq i < j \leq q\},$$

where $\mathsf{ind}^{(d)}_{\mathsf{diff}}(a, b)$ writes its inputs in base d as $a = (a[1], \ldots, a[m])$ and $b = (b[1], \ldots, b[m])$, and outputs $(k, |b[k] - a[k]|)$, where k is the smallest index such that $b[k] \neq a[k]$. If there is not such index (i.e. $a = b$) then it outputs $(m+1, 0)$.

Intuitively, this leakage outputs the index of the first base-d digit where each pair of messages differ, and additionally outputs the absolute difference in that digit. (When $d = 2$ the additional output is trivial, since it is always 1.)

We will show how to carry out the same strategy used in the proof of Theorem 2. Here we denote $m^* = m/\log d - 1$, and we will prove a version of Lemma 5 for a different nested sequence of pairs of messages $(i^L_j, i^R_j)^{m^*}_{j=1}$ that we define as follows:

$$i^L_j = 0, \quad i^R_j = 0^{m^*-j}||1||(d-1)^j$$

And we define the pair $\hat{i}^L_j, \hat{i}^R_j$ as:

$$\hat{i}^L_j = 0^{m^*-j}||0||(d-1)^j, \quad \hat{i}^R_j = 0^{m^*-j}||1||0^j$$

According to the leakage profile, we have $(\mathcal{E}_K(i^L_j), \mathcal{E}_K(i^R_j))$ and $(\mathcal{E}_K(\hat{i}^L_j), \mathcal{E}_K(\hat{i}^R_j))$ are statistical indistinguishable. Using the sequence $(i^L_j, i^R_j)^{m^*}_{j=1}$, the rest of the proof of Lemma 5 can be carried out. Finally, the rest of the proof of Theorem 2 can be applied exactly as before. Hence we have the lower bound:

$$n \geq \lambda(m/\log d) - (m/\log d)\log(m/\log d)$$

referring to d-ary CLWW is also almost optimal.

References

1. R. Agrawal, J. Kiernan, R. Srikant, and Y. 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. ACM, 2004.
2. A. Arasu, S. Blanas, K. Eguro, R. Kaushik, D. Kossmann, R. Ramamurthy, and R. Venkatesan. Orthogonal security with cipherbase. In *6th Biennial Conference on Innovative Data Systems Research (CIDR'13)*, January 2013.

3. A. Boldyreva, N. Chenette, Y. Lee, and A. O'Neill. Order-preserving symmetric encryption. In A. Joux, editor, *Advances in Cryptology – EUROCRYPT 2009*, volume 5479 of *Lecture Notes in Computer Science*, pages 224–241, Cologne, Germany, Apr. 26–30, 2009. Springer, Heidelberg, Germany.
4. A. Boldyreva, N. Chenette, and A. O'Neill. Order-preserving encryption revisited: Improved security analysis and alternative solutions. In P. Rogaway, editor, *Advances in Cryptology – CRYPTO 2011*, volume 6841 of *Lecture Notes in Computer Science*, pages 578–595, Santa Barbara, CA, USA, Aug. 14–18, 2011. Springer, Heidelberg, Germany.
5. D. Boneh, K. Lewi, M. Raykova, A. Sahai, M. Zhandry, and J. Zimmerman. Semantically secure order-revealing encryption: Multi-input functional encryption without obfuscation. In E. Oswald and M. Fischlin, editors, *Advances in Cryptology – EUROCRYPT 2015, Part II*, volume 9057 of *Lecture Notes in Computer Science*, pages 563–594, Sofia, Bulgaria, Apr. 26–30, 2015. Springer, Heidelberg, Germany.
6. N. Chenette, K. Lewi, S. A. Weis, and D. J. Wu. Practical order-revealing encryption with limited leakage. In *FSE*, 2016.
7. T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.
8. F. B. Durak, T. M. DuBuisson, and D. Cash. What else is revealed by order-revealing encryption? In *ACM CCS*, 2016. To appear.
9. S. Garg, C. Gentry, and S. Halevi. Candidate multilinear maps from ideal lattices. In T. Johansson and P. Q. Nguyen, editors, *Advances in Cryptology – EUROCRYPT 2013*, volume 7881 of *Lecture Notes in Computer Science*, pages 1–17, Athens, Greece, May 26–30, 2013. Springer, Heidelberg, Germany.
10. S. Garg, C. Gentry, S. Halevi, M. Raykova, A. Sahai, and B. Waters. Candidate indistinguishability obfuscation and functional encryption for all circuits. In *54th Annual Symposium on Foundations of Computer Science*, pages 40–49, Berkeley, CA, USA, Oct. 26–29, 2013. IEEE Computer Society Press.
11. P. Grubbs, K. Sekniqi, V. Bindschaedler, M. Naveed, and T. Ristenpart. Leakage-abuse attacks against order-revealing encryption. Cryptology ePrint Archive, Report 2016/895, 2016. `http://eprint.iacr.org/2016/895`.
12. W. Lu, A. L. Varna, and M. Wu. Security analysis for privacy preserving search of multimedia. In *2010 IEEE International Conference on Image Processing*, pages 2093–2096, Sept 2010.
13. R. A. Popa, C. M. S. Redfield, N. Zeldovich, and H. Balakrishnan. Cryptdb: protecting confidentiality with encrypted query processing. In *Proceedings of the 23rd ACM Symposium on Operating Systems Principles 2011, SOSP 2011, Cascais, Portugal, October 23-26, 2011*, pages 85–100, 2011.
14. C. Wang, N. Cao, J. Li, K. Ren, and W. Lou. Secure ranked keyword search over encrypted cloud data. In *Distributed Computing Systems (ICDCS), 2010 IEEE 30th International Conference on*, pages 253–262, June 2010.