On Hiding a Plaintext Length by Preencryption[⋆]

Cihangir Tezcan and Serge Vaudenay

EPFL
CH-1015 Lausanne, Switzerland
http://lasecwww.epfl.ch

Abstract. It is a well known fact that encryption schemes cannot hide a plaintext length when it is unbounded. We thus admit that an approximation of it may leak and we focus on hiding its precise value. Some standards such as TLS or SSH offer to do it by applying some pad-then-encrypt techniques. In this study, we investigate the information leakage when these techniques are used. We define the notion of padding scheme and its associated security. We show that when a padding length is uniformly distributed, the scheme is nearly optimal. We also show that the insecurity degrades linearly with the padding length.

1 Introduction

Although an encryption process makes a plaintext unreadable to adversaries, the resulting ciphertext may still leak some information. Practically, we can always distinguish an encrypted SMS message from an encrypted HD video stream. Namely, the length of a plaintext may give some information away and it can often be deduced from the ciphertext. For instance, the lengths of a plaintext and the corresponding ciphertext are identical or differ by a small number of bits when the encryption is done by a stream or a block cipher. One way of hiding the plaintext size is to use *random padding* before the encryption which appends a padding of random length in $\{1,2,\ldots,B\}$. In this work, we investigate the information leakage when a random padding is used.

Let us consider a symmetric encryption system in which encryption under a key K is denoted by Enc_K and decryption is denoted by Dec_K. In the Shannon model [10], the plaintext and the key are defined by independent random variables X and K. Perfect secrecy is defined by the statistical independence of X and $Y = \mathsf{Enc}_K(X)$. If this property is satisfied, we can easily see that the plaintext domain must be finite: if y is a possible value for Y, then $p = \Pr[Y = y]$ is positive. For any possible value x for X, we must have $\Pr[\mathsf{Enc}_K(x) = y] = p$ due to perfect secrecy. Since $\mathsf{Enc}_K(x) = y$ implies $\mathsf{Dec}_K(y) = x$, $p \leq \Pr[\mathsf{Dec}_K(y) = x]$.

[⋆] The present version corrects a few typos from the published one.

By summing over all possible x's, we deduce that the number of such x is bounded by $\frac{1}{p}$, which is a finite number.[1]

This impossibility result extends to weaker security notions. In [3,4], Chor and Kushilevitz consider α-weak security, given $\alpha \geq 1$, which states that for all possible x_1, x_2, y, we have

$$\frac{1}{\alpha}\Pr[Y = y|X = x_2] \leq \Pr[Y = y|X = x_1] \leq \alpha \Pr[Y = y|X = x_2]$$

Perfect secrecy corresponds to the $\alpha = 1$ case. Encryption over a countably infinite domain cannot be α-weak secure for any α: if x_2 and y are possible simultaneous values for X and Y, $p = \frac{1}{\alpha}\Pr[Y = y|X = x_2]$ is positive and we have $\Pr[Y = y|X = x_1] \geq p$ for all possible x_1. So, $\Pr[\mathsf{Dec}_K(y) = x_1] \geq p$ and the number of possible plaintexts is bounded by $\frac{1}{p}$.

In [9], Phan and Vaudenay consider ε-statistically extended *indistinguishability under one-time encryption* (extended IND-OTE game), given $\varepsilon < 1$, which means

$$\frac{1}{2}\sum_y |\Pr[Y = y|X = x_1] - \Pr[Y = y|X = x_2]| \leq \varepsilon$$

for all possible plaintexts x_1 and x_2. Again, secure (in this sense) encryption over an infinite (countable) domain is impossible.

A public-key cryptosystem is nothing but an encryption scheme in which Enc_K can be described by using public values. So, the above impossibility results also apply to public-key cryptography.

A standard security notion for encryption is the IND-CPA security (*indistinguishability under chosen plaintext attacks*) in which an adversary can make some chosen plaintext encryptions and tries to get an advantage for distinguishing the encryption of either x_1 or x_2, two plaintexts *of same length* selected by himself. For public-key encryption, the adversary makes the encryption himself by using the public key so IND-CPA and IND-OTE notions are equivalent. For symmetric encryption, he must be provided access to an encryption oracle. In the IND-OTE game, there is no such access so there may be a gap between IND-CPA and IND-OTE notions. Still, these notions impose that x_1 and x_2 have the same length so they offer no guarantee about keeping the plaintext length secret. We call *extended* IND-OTE game (E-IND-OTE) the notion where the restriction that x_1 and x_2 have the same length is relaxed.

[1] Actually, this proof only holds for countable sets. More generally, we should define our properties with non-discrete probabilities to be able to consider uncountably infinite sets. In theory, we *could* achieve perfect secrecy over uncountably infinite sets. However, the encryption algorithm will no longer be polynomially bounded on classical computational models. So, we only consider countable sets in the present paper. We could probably reopen this case when considering encryption over a space of quantum states. (See [9] for more discussions.)

The Phan-Vaudenay result says that if all adversaries in the $\mathsf{E\text{-}IND\text{-}OTE}$ game have their advantage bounded by a given $\varepsilon < 1$, then the encryption domain must be finite. Ideally, we would like to design secure encryption schemes over an infinite set. Practically, we could live with encryption domains which are finite but large enough. Indeed, we can assume that the set of bitstrings of length bounded by a few petabytes is a *virtually infinite* set. So, we could design an encryption scheme over this domain with a pretty good security. However, for efficiency reasons, we would not like that the encryption of a very small plaintext (say a few kilobytes) would lead to a ciphertext of one petabyte. Therefore, we should consider encryption schemes which are *somehow* length-preserving but also length hiding. To make it possible, we relax the $\mathsf{E\text{-}IND\text{-}OTE}$ security notion and consider the Δ-$\mathsf{IND\text{-}OTE}$ game in which the submitted plaintexts have a length difference bounded by Δ. The $\mathsf{IND\text{-}OTE}$ (resp. $\mathsf{E\text{-}IND\text{-}OTE}$) notions correspond to $\Delta = 0$ (resp. $\Delta = +\infty$).

In the sequel we consider encryption schemes defined by

$$\mathsf{Enc}(x) = \mathsf{Enc}_0(x \| \mathsf{pad}(x))$$

where Enc_0 is a length-preserving $\mathsf{IND\text{-}OTE}$-secure scheme and pad is a probabilistic padding scheme. That is, pad generates a postfix-free random string which can be extracted after decryption. Typically, $\mathsf{pad}(x)$ is a random bitstring whose length N is a random variable. This kind of construction is proposed e.g. in TLS [6] or SSH [12]. For instance, here is a quote from [6]:

> Padding that is added to force the length of the plaintext to be an integral multiple of the block cipher's block length. The padding MAY be any length up to 255 bytes, as long as it results in the TLSCiphertext.length being an integral multiple of the block length. Lengths longer than necessary might be desirable to frustrate attacks on a protocol that are based on analysis of the lengths of exchanged messages.

This suggests that we could arbitrarily pad up to $B = 32$ (resp. $B = 16$) blocks of data to hide the exact length of a plaintext, when the block cipher uses blocks of 64 bits (resp. 128 bits).

More generally, we consider preencryption schemes which are not necessarily of form $x \| \mathsf{pad}(x)$. We may consider several assumptions:

- (uniformity) the distribution of the length overhead between x and $\mathsf{Enc}(x)$ is fixed (it does not depend on x)
- (almost length-preserving property) the length overhead is bounded by B

Given B and Δ, our aim is to find the best distribution N to achieve optimal Δ-$\mathsf{IND\text{-}OTE}$ security.

Related work. Padding often serves another purpose. Namely, it is used to fill incomplete blocks to encrypt a plaintext using a block cipher. Our notion of preencryption scheme is similar to the notion of *encoding* by Paterson and Watson [8] who consider several practical schemes. They analyze the security of the pad-then-encrypt scheme in a practical case where the original encryption scheme is a block cipher in CTR mode. This follows some other work in which they identified a terrible interaction between the padding scheme and the decryption algorithm in CBC mode [1]. Some other padding schemes leading to decryption attacks have been identified (see e.g. [2,5,7,11]).

Our results. We first formalize in Section 2 the notion of preencryption scheme and its associated Δ-IND security notion. We formalize the notion of preencryption by padding (or the pad-then-encrypt technique). When Enc_0 is length-preserving, we show that Δ-IND-security is necessary and sufficient to make Enc Δ-IND-OTE secure.

Then, we show in Section 3 that there is always an adversary with advantage nearly $\frac{\Delta}{B}$. That is, the insecurity degrades linearly with the padding length B. This main result happens to have a very simple proof by using a diagonal argument.

We observe that a padding scheme making padding lengths uniformly distributed makes the above adversary nearly the best one. So, this preencryption scheme is nearly optimal.

In Section 4, we further precisely study the optimal padding scheme in the uniform case for $\Delta = 2$.

2 Preliminaries

In what follows we consider an alphabet Z. This can be a Boolean alphabet, or the set of bytes, or a set of blocks. We denote by Z^* the set of finite sequences of elements taken from Z. The length of an element $x \in Z^*$ is denoted by $|x|$. For $x, x' \in Z^*$, we denote by $x\|x'$ the concatenation of x and x'.

In this paper we adopt exact security notions. We can easily translate to asymptotic security by introducing security parameters in the definition of encryption schemes.

2.1 Encryption Scheme

Definition 1. *An* encryption scheme *is defined by*

- *a plaintext domain which is a subset of Z^**
- *an algorithm to generate a key K*

- *a (probabilistic) encryption algorithm* Enc *taking a key and a plaintext as input and producing a ciphertext*
- *a (deterministic) decryption algorithm* Dec *taking a key and a ciphertext as input and producing a plaintext*

The correctness property of an encryption scheme states that if we generate a key K by the key generation algorithm, if we take a plaintext x in the plaintext domain, and if we compute $\mathsf{Dec}_K(\mathsf{Enc}_K(x))$ then we obtain x with probability 1.

*We say that an encryption scheme is B-*almost length preserving *if*

$$||\mathsf{Enc}_K(x)| - |x|| \leq B$$

with probability 1 for all x. It is length-preserving if it is 0-almost length preserving.

*We say that an encryption scheme t-*fully leaks the plaintext length *if there exists an algorithm f such that for all x in the plaintext domain, $f(\mathsf{Enc}_K(x)) = |x|$ with probability 1 within a complexity at most t.*

For instance, a length-preserving encryption scheme fully leaks the plaintext length by $f(y) = |y|$.

For symmetric encryption, the key generation algorithm simply picks a key in a given key space, following the uniform distribution. For public-key encryption, the key can be split in a public part and a private part. The encryption algorithm only use the public part. What follows applies to both cases.

We define the Δ-IND-OTE security notion as follows:

Definition 2. *We consider the following game between an adversary $\mathcal{A}$ and a challenger. Firstly, the challenger generates a key K using the key generation algorithm. In the case of a public key cryptosystem, it reveals the public part K_p of K to the adversary. The adversary can do some computations and then submits two plaintexts $\mathcal{A}(K_p;\rho) = (x_0, x_1)$ in the plaintext domain such that*

$$||x_0| - |x_1|| \leq \Delta$$

by using some random coins ρ. The challenger flips a fair coin b, computes $Y = \mathsf{Enc}_K(x_b)$ and reveals Y. The adversary can then do some extra computations and yields a guess $\mathcal{A}(K_p, Y;\rho) = b'$. The adversary succeeds if $b = b'$. His advantage is $\Pr[b = b'] - \frac{1}{2}$. $\mathcal{A}$ has a complexity bounded by t if for any K_p, Y, and ρ, the total running time of $\mathcal{A}(K_p;\rho)$ and $\mathcal{A}(K_p, Y;\rho)$ is bounded by t. We say that the encryption scheme is Δ-IND-$\mathsf{OTE}(t,\varepsilon)$-secure if for all adversary with time complexity limited by t, the advantage is at most ε.

Δ-IND-OTE ***Game:***

```
1: Challenger generates $K$ and discloses its public part $K_p$
```

```
2: Adversary selects plaintexts $x_0$ and $x_1$ where $||x_0|-|x_1|| \leq \Delta$
3: Challenger flips a coin $b$, computes $\mathsf{Enc}_K(x_b) = Y$ and gives $Y$ to the adversary
4: Adversary guesses $b'$ and wins if $b' = b$
```

$\mathsf{IND\text{-}OTE}$ security corresponds to the $\Delta = 0$ case. We also consider $\mathsf{E\text{-}IND\text{-}OTE}$ security defined by the $\Delta = +\infty$ case.

2.2 Preencryption Schemes

Definition 3. *Given two plaintext domains $\mathcal{X}$ and $\mathcal{X}^0$, a preencryption scheme from $\mathcal{X}$ to $\mathcal{X}^0$ is a pair of algorithms*

- *a (probabilistic) algorithm* pre *such that for all $x \in \mathcal{X}$,* $\mathsf{pre}(x) \in \mathcal{X}^0$ *with probability 1*
- *a (deterministic) algorithm* $\mathsf{Extract}$

The correctness property of a preencryption scheme states that for all $x \in \mathcal{X}$,

$$\mathsf{Extract}(\mathsf{pre}(x)) = x$$

with probability 1.

*We say that a preencryption scheme is B-*almost length preserving *if*

$$||\mathsf{pre}(x)| - |x|| \leq B$$

with probability 1 for all x. We say that a preencryption scheme is length-increasing *if* $|\mathsf{pre}(x)| \geq |x|$ *with probability 1 for all x. We say that a preencryption scheme is* strictly length-increasing *if* $|\mathsf{pre}(x)| > |x|$ *with probability 1 for all x.*

Definition 4. *A preencryption scheme is* Δ-IND (t,ε)*-secure if for all adversary $\mathcal{A}$ with time complexity limited by t, the advantage in the following game is at most ε. The advantage is defined as* $\Pr[b = b'] - \frac{1}{2}$.

```
Δ-IND Game:
1: Adversary selects plaintexts $x_0$ and $x_1$ where $||x_0|-|x_1|| \leq \Delta$
2: Challenger flips a coin $b$, computes $|\mathsf{pre}(x_b)| = L$ and gives $L$ to the adversary
3: Adversary guesses $b'$ and wins if $b' = b$
```

$\mathcal{A}$ has a complexity bounded by t if for any L and ρ, the total running time of $\mathcal{A}(;\rho)$ and $\mathcal{A}(L;\rho)$ is bounded by t.

Given a set of integers A, x_0 and x_1, we define a Δ-IND adversary $D_A(x_0,x_1)$ as the one selecting x_0 and x_1 then yielding $b' = 1$ if and only if $L \in A$. We define $\mathsf{Adv}_A(x_0,x_1)$ as the advantage of this adversary.

Lemma 5. *For any x_0 and x_1 we have*

$$\begin{aligned}\mathsf{Adv}_A(x_0,x_1) &= \frac{1}{2}\Pr[|\mathsf{pre}(x_1)| \in A] - \frac{1}{2}\Pr[|\mathsf{pre}(x_0)| \in A] \\ &= \frac{1}{2}\sum_{\ell\in A}(\Pr[|\mathsf{pre}(x_1)| = \ell] - \Pr[|\mathsf{pre}(x_0)| = \ell])\end{aligned}$$

Proof. We have

$$\begin{aligned}\mathsf{Adv}_A(x_0,x_1) &= \Pr[b=b'] - \frac{1}{2} \\ &= \frac{1}{2}\Pr[b'=1|b=1] + \frac{1}{2}\Pr[b'=0|b=0] - \frac{1}{2} \\ &= \frac{1}{2}\Pr[b'=1|b=1] - \frac{1}{2}\Pr[b'=1|b=0] \\ &= \frac{1}{2}\Pr[|\mathsf{pre}(x_1)| \in A] - \frac{1}{2}\Pr[|\mathsf{pre}(x_0)| \in A]\end{aligned}$$

and the other expression follows by separating the $\ell \in A$ cases. □

We define $\mathsf{Adv}(x_0,x_1)$ as the maximal advantage for (computationally unbounded) adversaries selecting x_0 and x_1.

Lemma 6. *For any x_0 and x_1 we have* $\mathsf{Adv}(x_0,x_1) = \mathsf{Adv}_A(x_0,x_1)$ *where*

$$A = \{\ell; \Pr[|\mathsf{pre}(x_1)| = \ell] > \Pr[|\mathsf{pre}(x_0)| = \ell]\}$$

Actually, $\mathsf{Adv}(x_0,x_1)$ is the statistical distance between the length of $\mathsf{pre}(x_0)$ and the length of $\mathsf{pre}(x_1)$.

Proof. Since we consider unbounded adversaries, an optimal one using x_0 and x_1 can be assumed to be of form $D_{A'}(x_0,x_1)$ without loss of generality. By Lemma 5 we clearly have $\mathsf{Adv}_{A'}(x_0,x_1) \leq \mathsf{Adv}_A(x_0,x_1)$. So, $A' = A$ maximizes $\mathsf{Adv}_{A'}(x_0,x_1)$ and we obtain $\mathsf{Adv}(x_0,x_1) = \mathsf{Adv}_A(x_0,x_1)$. □

Given an encryption scheme

$$C^0 = (\mathcal{X}^0, \mathsf{Gen}^0, \mathsf{Enc}^0, \mathsf{Dec}^0)$$

and a preencryption scheme $P = (\mathsf{pre}, \mathsf{Extract})$ from $\mathcal{X}$ to $\mathcal{X}^0$ we define the encryption scheme

$$C = (\mathcal{X}, \mathsf{Gen}, \mathsf{Enc}, \mathsf{Dec})$$

by $\mathsf{Gen} = \mathsf{Gen}^0$,

$$\mathsf{Enc}_K(x) = \mathsf{Enc}^0_K(\mathsf{pre}(x))$$

and

$$\mathsf{Dec}_K(y) = \mathsf{Extract}(\mathsf{Dec}^0_K(y))$$

Clearly, this defines an encryption scheme. If the preencryption scheme is B-almost length-preserving and the encryption scheme C^0 is length-preserving, then the encryption scheme C is B-almost length preserving.

Theorem 7. *We assume there exist a constant t_S and a sampling algorithm $S(1^L)$ to pick a random element of $\mathcal{X}^0$ of length L with complexity at most t_S for any $L \in \{|x|; x \in \mathcal{X}^0\}$. There exists a (small) constant c such that for any C, C_0, t, t_P, if C^0 is a $\mathsf{IND\text{-}OTE}(t+t_S+t_P+c, \varepsilon^0)$-secure encryption scheme and if P is a $\Delta\text{-}\mathsf{IND}(t+t_S+t_P+c, \varepsilon^1)$-secure preencryption scheme and* pre *can be computed within a complexity bounded by t_P, then C is a $\Delta\text{-}\mathsf{IND\text{-}OTE}(t, 2\varepsilon^0+\varepsilon^1)$-secure encryption scheme.*

When C^0 t_0-fully leaks the plaintext length, if C is $\Delta\text{-}\mathsf{IND\text{-}OTE}(t+t_0+c, \varepsilon)$-secure then P is $\Delta\text{-}\mathsf{IND}(t, \varepsilon)$-secure.

So, for an IND-OTE-secure encryption C^0 which fully leaks the plaintext length, the Δ-IND security of P is necessary and sufficient to have C Δ-IND-OTE-secure.

Proof. Let $\mathcal{A}$ be a Δ-IND-OTE adversary for C which has a time complexity bounded by t. We want to prove that its advantage is less than $2\varepsilon^0+\varepsilon^1$.

We define the following adversary $\mathcal{A}'$.

```
receive (public) key material
simulate $\mathcal{A}$ to get $x_0$ and $x_1$
flip a fair coin $b$
compute $x'_0 = \mathsf{pre}(x_b)$
pick a random $x'_1 = S(1^{|x'_0|})$ in $\mathcal{X}^0$
submit $x'_0$ and $x'_1$ and receive $Y$
continue the simulation of $\mathcal{A}$ with $Y$ to get $b'$
output 1 if $b = b'$ and 0 otherwise
```

The complexity of this adversary is bounded by $t+t_S+t_P+c$ where c is the small overhead complexity beside the simulation of $\mathcal{A}$, the sampling of S, and the computation of $\mathsf{pre}(x_b)$.

Let Γ be the experiment corresponding to the IND-OTE game of $\mathcal{A}'$ against C^0 when x'_0 is selected by the challenger. So, Γ yields 1 if and only if $\mathcal{A}$ yields $b' = b$ on input $Y = \mathsf{Enc}(\mathsf{pre}(x_0))$. That is, the advantage of $\mathcal{A}$ is $\Pr[\Gamma \to 1] - \frac{1}{2}$. Therefore, to bound the advantage of $\mathcal{A}$, we just need to prove that $\Pr[\Gamma \to 1] \leq \frac{1}{2} + 2\varepsilon^0 + \varepsilon^1$.

Let Γ' be the experiment corresponding to the IND-OTE game of $\mathcal{A}'$ against C^0 when x'_1 is selected by the challenger. $\mathcal{A}'$ is an IND-OTE adversary for C^0

with advantage $\frac{1}{2}(\Pr[\Gamma' \to 1] - \Pr[\Gamma \to 1])$. Due to the IND-OTE-security of C^0, we have $|\Pr[\Gamma \to 1] - \Pr[\Gamma' \to 1]| \leq 2\varepsilon^0$.

Clearly, Γ' is equivalent to the following:

```
generate a key
simulate $\mathcal{A}$ to get $x_0$ and $x_1$
flip a fair coin $b$
compute $L = |\mathsf{pre}(x_b)|$
pick $X = S(1^L)$
compute $Y = \mathsf{Enc}(X)$
continue the simulation of $\mathcal{A}$ with $Y$ to get $b'$
output 1 if $b = b'$ and 0 otherwise
```

This defines a Δ-IND adversary for P. So, $\Pr[\Gamma' \to 1] \leq \frac{1}{2} + \varepsilon^1$.

We deduce that $\Pr[\Gamma \to 1] \leq \frac{1}{2} + 2\varepsilon^0 + \varepsilon^1$.

For the second part of the theorem, we now let $\mathcal{A}$ be a Δ-IND adversary for P of complexity bounded by t and we want to bound its advantage. Since C^0 fully leaks the plaintext length, there is a function f to compute the plaintext length from the ciphertext. We define the following adversary:

```
get key material from a challenger
simulate $\mathcal{A}$ to get $x_0$ and $x_1$
submit $x_0$ and $x_1$ to the challenger and get ciphertext $Y$
compute $L = f(Y)$
continue the simulation of $\mathcal{A}$ with $L$ to get $b'$
yield $b'$
```

Clearly, this is a Δ-IND-OTE adversary for C whose advantage is exactly the advantage of $\mathcal{A}$. Assume that its complexity is bounded by $t + t_0 + c$. Since C is Δ-IND-OTE $(t + t_0 + c, \varepsilon)$-secure, this advantage is bounded by ε. □

2.3 Pad-then-Encrypt Scheme

Definition 8. *A C subset of Z^* is postfix-free if*

$$\forall s \in Z^* \quad \forall x, y \in C \quad s\|x = y \Longrightarrow x = y$$

We observe that if the empty string belongs to C then no other string is in C. Furthermore, there exists a function Extract such that for all $s \in \mathcal{X}$ and for all $x \in C$, we have

$$\mathsf{Extract}(s\|x) = s$$

with probability 1. In what follows we consider a postfix-free set such that this function can be efficiently implemented.

Definition 9. *Given $\mathcal{X}^0 \subseteq Z^*$ and a postfix-free set C, a C-padding scheme on $\mathcal{X}^0$ is a probabilistic algorithm taking an element x of $\mathcal{X}^0$ as input and producing an element* $\mathsf{pad}(x)$ *of C as an output. We say that the padding scheme is* uniform *if the distribution of* $\mathsf{pad}(x)$ *does not depend on x.*

A padding scheme defines the preencryption scheme

$$\mathsf{pre}(x) = x \| \mathsf{pad}(x)$$

We note that preencryption schemes made out from a padding scheme are all length-increasing. Except in the constant 0-padding case, they are even *strictly* length increasing.

Example 10. We consider the padding scheme defined by the parameter B as follows: given x, we simply pick a sequence $100\cdots0$ of length N which is uniformly distributed in $\{1,\ldots,B\}$. This padding scheme is B-almost length preserving, strictly length-increasing, and uniform. By Lemma 5 and 6, we obtain that $\mathsf{Adv}(x_0,x_1) = \frac{||x_1|-|x_0||}{2B}$. So, this preencryption scheme is Δ-$\mathsf{IND}\left(t, \frac{\Delta}{2B}\right)$-secure for all Δ and any t.

In what follows we show that this scheme is nearly optimal.

To make a pad-then-encrypt construction secure with Δ large, we shall find a secure padding scheme for this Δ. A trivial solution consists of making sure that $x\|\mathsf{pad}(x)$ has a constant length no matter the plaintext x. To make it possible, this length must be at least the maximal length of a plaintext. This solution is clearly impractical. We shall rather concentrate on Δ small. So, we do not fully hide the length of plaintexts but rather their exact value.

3 Maximal Security of the Pad-then-Encrypt Scheme

In this section we consider lower bounds for the best advantage of an adversary against a preencryption scheme. We consider the case where the plaintext space is large and dense enough so that we can make sequences of plaintexts such that the length of two consecutive ones differ by Δ.

Definition 11. *We say that a sequence $(x_0,\ldots,x_n)$ of Z^* elements is a* Δ-chain *if for every $i = 0,\ldots,n-1$, we have $|x_{i+1}| - |x_i| = \Delta$. We say that this sequence* represents *a length ℓ if $|x_0| \leq \ell \leq |x_n|$. We say that a subset $\mathcal{X}$ of Z^* is* Δ-dense *if for any $x,y \in \mathcal{X}$, there exists a Δ-chain in $\mathcal{X}$ which represents $|x|$ and $|y|$. We say that $\mathcal{X}$ is B-*large *if there exists $x,y \in \mathcal{X}$ such that $|x| - |y| \geq B$.*

Theorem 12. *Let P be a B-almost length-preserving preencryption scheme and Δ be an integer. We assume that the input domain of P is Δ-dense and $(2B+\Delta)$-large. Then, there exists an adversary in the* Δ-IND *game with advantage at least* $1/\left(2\left\lfloor\frac{2B}{\Delta}\right\rfloor+2\right)$.

If P is length-increasing and B-almost length-preserving over a domain which is Δ-dense and $(B+\Delta)$-large, then there exists an adversary with advantage at least $1/\left(2\left\lfloor\frac{B}{\Delta}\right\rfloor+2\right)$.

Proof. Let $n=\left\lfloor\frac{cB}{\Delta}\right\rfloor+1$ with $c=1$ for length-increasing preencryption schemes and $c=2$ otherwise. Since the domain is $(cB+\Delta)$-large and Δ-dense, we can construct a Δ-chain of $n+1$ elements $x_0,x_1,\ldots,x_n$. We have $|x_{i+1}|=|x_i|+\Delta$ for $i=0,1,\ldots,n-1$. So, $|x_i|=|x_0|+i\Delta$ for $i=0,1,\ldots,n$. Let

$$s_i=\Pr[|\mathsf{pre}(x_i)|\leq B+|x_0|]$$

which is the probability that the preencrypted version of x_i has an overhead length bounded by $B+|x_0|-|x_i|=B-i\Delta$. Clearly, $s_0=1$ since P is B-almost length preserving, and $s_n=0$ since $B-n\Delta<(1-c)B$.

So, $\sum_{i=0}^{n-1}(s_i-s_{i+1})=1$. Hence, there must exist some i such that $s_i-s_{i+1}\geq\frac{1}{n}$. Let A be the set of all integers up to $B+|x_0|$. We have $\Pr[|\mathsf{pre}(x_i)|\in A]=s_i$. We deduce that $\mathsf{Adv}_A(x_i,x_{i+1})\geq\frac{1}{2n}$: there is an adversary with an advantage larger than $\frac{1}{2n}$. □

Remark 13. Example 10 shows a simple B-almost length-preserving scheme which is Δ-IND $\left(t,\frac{\Delta}{2B}\right)$-secure. So, the optimal security which is achievable is between $\frac{\Delta}{2B}$ and $\frac{1}{2\left\lceil\frac{B}{\Delta}\right\rceil}$. In particular, when Δ divides B, the scheme in Example 10 is optimal.

Theorem 12 can be generalized to preencryption schemes which are unbounded, but with finite expected overhead length. In practice, we would like to have a guarantee that a preencryption overhead is not too long on average, so this is a pretty reasonable assumption.

Theorem 14. *Let P be a length-increasing preencryption scheme and Δ be an integer. We assume that the input domain of P is Δ-dense and $(2B)$-large. We assume that for all x, we have $|E(|\mathsf{pre}(x)|)-|x||\leq B$. There exists an adversary in the* Δ-IND *game with advantage at least* $1/\left(4\left\lceil\frac{2B}{\Delta}\right\rceil\right)$.

Proof. We apply the same proof method as in Theorem 12. We define $n=\left\lceil\frac{\alpha B}{\Delta}\right\rceil$ and

$$s_i=\Pr[|\mathsf{pre}(x_i)|<\alpha B+|x_0|]=\Pr[|\mathsf{pre}(x_i)|-|x_i|<\alpha B-i\Delta]$$

for α such as the scheme is (αB)-large. We have $s_0\geq1-\frac{1}{\alpha}$ since $E(|\mathsf{pre}(x_0)|)-|x_0|\leq B$ and $s_n=0$ since the scheme is length-increasing. So, there is some i

leading us to $\mathsf{Adv}_A(x_i, x_{i+1}) \geq \frac{1}{2n}\left(1 - \frac{1}{\alpha}\right)$. We can just take $\alpha = 2$ and conclude. □

4 Uniform Padding Schemes

In this section, we consider a uniform padding scheme. We let N be a random variable following the distribution of $|\mathsf{pad}(x)|$. We assume that $\Pr[N = 0] = 0$: the padding scheme is strictly length-increasing. Since the scheme is uniform, the distribution does not depend on x. In notations, we further replace plaintexts x_0 and x_1 by their lengths a and b where $b \geq a$ without loss of generality.

Lemma 15. *We have* $\Pr[N \leq b - a] \leq 2 \cdot \mathsf{Adv}(a, b)$ *and equality holds if and only if* $\Pr[N = x + b - a] \leq \Pr[N = x]$ *for all* $x > 0$.

Proof. Let $\varepsilon = \mathsf{Adv}(a, b)$. Due to Lemma 5 and 6, we have

$$\begin{aligned}\varepsilon &= \frac{1}{2} \sum_{\ell:\Pr[N=\ell-a]\geq\Pr[N=\ell-b]} (\Pr[N = \ell - a] - \Pr[N = \ell - b]) \\ &\geq \frac{1}{2} \sum_{\ell:\ell\leq b} \Pr[N = \ell - a] \\ &= \frac{1}{2} \Pr[N \leq b - a]\end{aligned}$$

and equality holds if and only if $\Pr[N = x + b - a] \leq \Pr[N = x]$ for all $x > 0$. □

Theorem 16. *Consider a uniform strictly length-increasing padding scheme with the above notations. We assume that it is B-almost length-preserving. If* $b - a = \Delta$ *and B is divisible by* Δ*, then* $\mathsf{Adv}(a, b) \geq \frac{\Delta}{2B}$ *and equality holds if and only if* $\Pr[N \leq b - a] = \frac{\Delta}{B}$ *and* $\Pr[N = i]$ *is periodic over* $[1, \ldots, B]$ *with period* Δ.

Proof. Let $\varepsilon = \mathsf{Adv}(a, b)$.

Case 1: Assume $\Pr[N \leq \Delta] > \frac{\Delta}{B}$. Due to Lemma 15, we have $\varepsilon \geq \frac{1}{2}\Pr[N \leq \Delta] > \frac{\Delta}{2B}$.

Case 2: Assume $\Pr[N \leq \Delta] = \frac{\Delta}{B}$. If there exists an integer $j > a + \Delta$ with $\Pr[N = j - a] > \Pr[N = j - b]$, then $A = \{a + 1, a + 2, \ldots, a + \Delta, j\}$ makes

$$\varepsilon \geq \mathsf{Adv}_A(a, b) = \frac{1}{2}(\Pr[N \leq \Delta] + \Pr[N = j - a] - \Pr[N = j - b]) > \frac{\Delta}{2B}$$

If no such j exists, then we have $\Pr[N = x+\Delta] \leq \Pr[N = x]$ for all $x > 0$. By Lemma 15, we obtain $\varepsilon = \frac{\Delta}{2B}$. Furthermore, we get $\Pr[j\Delta < N \leq (j+1)\Delta] \leq \frac{\Delta}{B}$ for all $j \geq 0$. Therefore, we have

$$1 = \sum_{i=1}^{B} \Pr[N = i] \leq \left\lceil \frac{B}{\Delta} \right\rceil \cdot \frac{\Delta}{B}$$

Since B is divisible by Δ, this inequality is in fact an equality. Thus, we cannot have $\Pr[N = x+\Delta] < \Pr[N = x]$ for any x. Hence, $\Pr[N = x+\Delta] = \Pr[N = x]$ for all $x \in [1, B-\Delta]$, and $\Pr[N = i]$ becomes periodic over $[1,\ldots,B]$ with period Δ.

Case 3: Assume $\Pr[N \leq \Delta] < \frac{\Delta}{B}$. Then $\frac{\Delta}{B} - \Pr[N \leq \Delta] = \delta$ for some $\delta > 0$. Since

$$\sum_{j=0}^{\lceil \frac{B}{\Delta} \rceil - 1} \Pr[j\Delta < N \leq (j+1)\Delta] = 1$$

$\Pr[0 < N \leq \Delta] = \frac{\Delta}{B} - \delta$, and Δ divides B, there must exist an integer $j > 0$ such that $\Pr[j\Delta < N \leq (j+1)\Delta] > \frac{\Delta}{B}$. Thus, if we set $A = \{a+1, a+2, \ldots, a+(j+1)\Delta\}$, we obtain

$$\varepsilon \geq \mathsf{Adv}_A(a,b) = \frac{1}{2}\left(\Pr[N \leq (j+1)\Delta] - \Pr[N \leq j\Delta]\right) > \frac{\Delta}{2B}$$

Thus in all cases $\varepsilon \geq \frac{\Delta}{2B}$ and equality holds if and only if $\Pr[N \leq \Delta] = \frac{\Delta}{B}$ and $\Pr[N = i]$ is periodic over $[1,\ldots,B]$ with period Δ. □

The following example shows that when B is not divisible by Δ, then ε can be less than $\frac{\Delta}{B}$.

Example 17. Let $b - a = \Delta = 2$ and $B = 5$. We define N as follows:

$$\Pr[N = 1] = \Pr[N = 3] = \Pr[N = 5] = 0.22 \qquad \Pr[N = 2] = \Pr[N = 4] = 0.17$$

Thus, the best advantage with a length difference of $\Delta = 2$ is $\varepsilon_2 = \frac{1}{2}(\Pr[N = 1] + \Pr[N = 2]) = 0.195$ which is less than $\frac{1}{5}$. However, for $\Delta = 1$, the best advantage is $\varepsilon_1 = \frac{1}{2}(\Pr[N = 1] + \Pr[N = 3] + \Pr[N = 5] - \Pr[N = 2] - \Pr[N = 4]) = 0.16$.

For $\Delta = 1$, B is divisible by Δ so Example 10 gives an optimal padding scheme. For $\Delta = 2$ and B even, it is the same. For $\Delta = 2$ and B odd, the optimal case is characterized as follows.

Theorem 18. *Consider a uniform strictly length-increasing padding scheme with the above notations. We assume that it is B-almost length-preserving. If B is odd, then*

$$\max_{b-a \leq 2} \mathsf{Adv}(a,b) \geq \frac{B}{B^2+1}$$

and an equality can be reached by a distribution taking alternate values on every length.

Proof. We first note that $\frac{B}{B^2+1} < \frac{1}{B}$ so we must find a better distribution than the uniform one from Example 10. We further note that $\frac{1}{2\lfloor\frac{B}{2}\rfloor+2} = \frac{1}{B+1} \leq \frac{B}{B^2+1}$ so the bound to be proven is consistent with the one from Theorem 12. Let

$$\begin{aligned}\varepsilon &= \max_{b-a\leq 2} \mathsf{Adv}(a,b)\\ \varepsilon_1 &= \max_{b-a=1} \mathsf{Adv}(a,b)\\ \varepsilon_2 &= \max_{b-a=2} \mathsf{Adv}(a,b)\end{aligned}$$

We have $\varepsilon = \max(\varepsilon_1, \varepsilon_2)$.

We let $\alpha = \frac{B-1}{B(B^2+1)}$, $\beta = \frac{B+1}{B(B^2+1)}$. We note that $\frac{2}{B} + \alpha - \beta = \frac{2B}{B^2+1}$. Furthermore, $\frac{B+1}{2}\alpha - \frac{B-1}{2}\beta = 0$. Let N_0 be a random variable defined by the distribution $\Pr[N_0 = i] = \frac{1}{B} + \alpha$ for i odd and $\Pr[N_0 = i] = \frac{1}{B} - \beta$ for i even. That is, the distribution of N_0 takes alternate values on every length. For $N = N_0$, by using Lemma 5 and Lemma 6, we obtain $\varepsilon_1 = \frac{1}{2}\left(\frac{1}{B} + \alpha + \frac{B-1}{2}(\alpha+\beta)\right) = \frac{B}{B^2+1}$ with the optimal set $A = \{a+1, a+3, \ldots, a+B\}$ and $\varepsilon_2 = \frac{1}{2}\left(\frac{2}{B} + \alpha - \beta\right) = \frac{B}{B^2+1}$ with the optimal set $A = \{a+1, a+2\}$. So, $\varepsilon = \frac{B}{B^2+1}$. We now want to prove that there is no distribution for N achieving a lower ε.

Let us assume that there is some $0 \leq i < B-1$ such that $\Pr[N \in \{i+1, i+2\}] > \Pr[N_0 \in \{i+1, i+2\}] = \frac{2B}{B^2+1}$. We take $A = \{a+1, a+2, \ldots, a+i+2\}$ and we obtain

$$\varepsilon \geq \mathsf{Adv}_A(a, a+2) = \frac{1}{2}\left(\Pr[N \leq i+2] - \Pr[N \leq i]\right) > \frac{B}{B^2+1}$$

which is not better than our above distribution. Hence, we now assume that $\Pr[N \in \{i+1, i+2\}] \leq \Pr[N_0 \in \{i+1, i+2\}]$ for $i = 0, \ldots, B-2$.

Let i be an odd integer. Since $\Pr[N \in \{u, u+1\}] \leq \Pr[N_0 \in \{u, u+1\}]$ for $u = 1, 3, \ldots, i-2, i+1, \ldots, B-3, B-1$, by summing all inequalities, we obtain $\Pr[N \neq i] \leq \Pr[N_0 \neq i]$. So,

$$\Pr[N = i] = 1 - \Pr[N \neq i] \geq 1 - \Pr[N_0 \neq i] = \Pr[N_0 = i]$$

for any odd i. Thus, $\Pr[N \text{ odd}] \geq \Pr[N_0 \text{ odd}]$.

Let now i be even. We have

$$\begin{aligned}\Pr[N = i] &= \Pr[N \in \{i, i+1\}] - \Pr[N = i+1]\\ &\leq \Pr[N_0 \in \{i, i+1\}] - \Pr[N_0 = i+1]\\ &= \Pr[N_0 = i]\end{aligned}$$

Thus, $\Pr[N \text{ even}] \leq \Pr[N_0 \text{ even}]$.

Finally, let $A = \{a+1, a+3, \ldots, B\}$. We have

$$\begin{aligned}\varepsilon \geq \mathsf{Adv}_A(a, a+1) &= \frac{1}{2}\left(\Pr[N \text{ odd}] - \Pr[N \text{ even}]\right) \\ &\geq \frac{1}{2}\left(\Pr[N_0 \text{ odd}] - \Pr[N_0 \text{ even}]\right) \\ &= \frac{B}{B^2+1}\end{aligned}$$

Therefore, we cannot have ε lower than $\frac{B}{B^2+1}$. □

Theorem 18 shows that when $b - a \leq 2$ and B is odd, the lower bound $\frac{1}{2\lfloor\frac{B}{2}\rfloor+2} = \frac{1}{B+1}$ for the maximum advantage is not achievable. Results of Theorem 12 and 18 for the case when $\Delta = 2$ and B is odd are provided in Table 1 for small values of B.

5 Conclusion

We have shown that a padding scheme adding strings with uniformly distributed length is nearly optimal and that its security is roughly $\frac{\Delta}{2B}$. The optimal scheme can be slightly better but still close to this bound. This shows that the price to pay for making ε-indistinguishable two plaintexts with a single bit of length difference (i.e. 1-$\mathsf{IND\text{-}OTE}(t, \varepsilon)$-security) is to append a padding of length $\frac{\varepsilon^{-1}}{2}$, which is impractical for the usual security levels we target for encryption (e.g. $\varepsilon = 2^{-80}$).

Acknowledgements. The authors would like to thank one of the reviewers for his/her lengthy and valuable comments.

References

1. M. R. Albrecht, G. J. Watson and K. G. Paterson. Plaintext Recovery Attacks Against SSH. In *IEEE Symposium on Security and Privacy*, Berkeley, CA, USA, pp. 16–26, IEEE, 2009.
2. B. Canvel, A. P. Hiltgen, S. Vaudenay, M. Vuagnoux. Password Interception in a SSL/TLS Channel. In *Advances in Cryptology CRYPTO'03*, Santa Barbara, California, U.S.A., Lecture Notes in Computer Science 2729, pp. 583–599, Springer-Verlag, 2003.
3. B. Chor, E. Kushilevitz. Secret Sharing over Infinite Domains (Extended Abstract). In *Advances in Cryptology CRYPTO'89*, Santa Barbara, California, U.S.A., Lecture Notes in Computer Science 435, pp. 299–306, Springer-Verlag, 1990.
4. B. Chor and E. Kushilevitz. Secret Sharing over Infinite Domains. *Journal of Cryptology*, vol. 6, pp. 87–95, 1993.

Table 1. Results of the Theorem 12 and 18 when $\Delta = 2$ and B is odd

B	Upper bound (Ex. 10)	Best Achievable (Th. 18)	Lower Bound (Th. 12)
3	0.333333333333333	0.3	0.25
5	0.2	0.192307692307692	0.166666666666667
7	0.142857142857143	0.14	0.125
9	0.111111111111111	0.109756097560976	0.1
11	0.0909090909090909	0.0901639344262295	0.0833333333333333
13	0.0769230769230769	0.0764705882352941	0.0714285714285714
15	0.0666666666666667	0.0663716814159292	0.0625
17	0.0588235294117647	0.0586206896551724	0.0555555555555556
19	0.0526315789473684	0.0524861878453039	0.05
21	0.0476190476190476	0.0475113122171946	0.0454545454545455
23	0.0434782608695652	0.0433962264150943	0.0416666666666667
25	0.04	0.0399361022364217	0.0384615384615385
27	0.037037037037037	0.036986301369863	0.0357142857142857
29	0.0344827586206897	0.0344418052256532	0.0333333333333333
31	0.032258064516129	0.0322245322245322	0.03125
33	0.0303030303030303	0.0302752293577982	0.0294117647058824
35	0.0285714285714286	0.0285481239804241	0.0277777777777778
37	0.027027027027027	0.027007299270073	0.0263157894736842
39	0.0256410256410256	0.0256241787122208	0.025
41	0.024390243902439	0.0243757431629013	0.0238095238095238
43	0.0232558139534884	0.0232432432432432	0.0227272727272727
45	0.0222222222222222	0.0222112537018756	0.0217391304347826
47	0.0212765957446809	0.0212669683257919	0.0208333333333333
49	0.0204081632653061	0.0203996669442132	0.02
51	0.0196078431372549	0.0196003074558032	0.0192307692307692
53	0.0188679245283019	0.0188612099644128	0.0185185185185185
55	0.0181818181818182	0.0181758096497026	0.0178571428571429
57	0.0175438596491228	0.0175384615384615	0.0172413793103448
59	0.0169491525423729	0.0169442848937392	0.0166666666666667
61	0.0163934426229508	0.0163890381515314	0.0161290322580645
63	0.0158730158730159	0.0158690176322418	0.015625
65	0.0153846153846154	0.0153809749171794	0.0151515151515152
67	0.0149253731343284	0.0149220489977728	0.0147058823529412
69	0.0144927536231884	0.0144897102057959	0.0142857142857143
71	0.0140845070422535	0.014081713605712	0.0138888888888889
73	0.0136986301369863	0.0136960600375235	0.0135135135135135
75	0.0133333333333333	0.0133309633842872	0.0131578947368421
77	0.012987012987013	0.0129848229342327	0.0128205128205128
79	0.0126582278481013	0.012656199935918	0.0125

5. J.-P. Degabriele, K. G. Paterson. Attacking the IPsec Standards in Encryption-only Configurations. In *IEEE Symposium on Security and Privacy*, Berkeley, CA, USA, pp. 335–349, IEEE, 2007.
6. T. Dierks, C. Rescola. The TLS Protocol Version 1.2. RFC 5246, standard tracks, the Internet Society, 2008.
7. K. G. Paterson, A. K. L. Yau. Cryptography in Theory and Practice: The Case of Encryption in IPsec. In *Advances in Cryptology EUROCRYPT'06*, St. Petersburg, Russia, Lecture Notes in Computer Science 4004, pp. 12–29, Springer-Verlag, 2006. 2006
8. K. G. Paterson, G. J. Watson. Plaintext-Dependent Decryption: A Formal Security Treatment of SSH-CTR. In *Advances in Cryptology EUROCRYPT'10*, French Riviera, France, Lecture Notes in Computer Science 6110, pp. 345–361, Springer-Verlag, 2010.
9. R. C.-W. Phan, S. Vaudenay. On the Impossibility of Strong Encryption over $\aleph_0$. In *International Workshop on Coding and Cryptology IWCC'09*, Zhangjiajie, China, Lecture Notes in Computer Science 5557, pp. 202–218, Springer-Verlag, 2009.
10. C. E. Shannon. Communication Theory of Secrecy Systems. *Bell system technical journal*, vol. 28, pp. 656–715, 1949.
11. S. Vaudenay. Security Flaws Induced by CBC Padding — Applications to SSL, IPSEC, WTLS... In *Advances in Cryptology EUROCRYPT'02*, Amsterdam, Netherlands, Lecture Notes in Computer Science 2332, pp. 534–545, Springer-Verlag, 2002.
12. T. Ylonen. The Secure Shell (SSH) Transport Layer Protocol. RFC 4253, standard tracks, the Internet Society, 2006.