The Curse of Small Domains:
New Attacks on Format-Preserving Encryption

Viet Tung Hoang[1], Stefano Tessaro[2], and Ni Trieu[3]

[1] Dept. of Computer Science, Florida State University, USA.
[2] Dept. of Computer Science, University of California Santa Barbara, USA.
[3] Dept. of Computer Science, Oregon State University, USA.

Abstract. Format-preserving encryption (FPE) produces ciphertexts which have the same format as the plaintexts. Building secure FPE is very challenging, and recent attacks (Bellare, Hoang, Tessaro, CCS '16; Durak and Vaudenay, CRYPTO '17) have highlighted security deficiencies in the recent NIST SP800-38G standard. This has left the question open of whether practical schemes with high security exist.

In this paper, we continue the investigation of attacks against FPE schemes. Our first contribution are new known-plaintext message recovery attacks against Feistel-based FPEs (such as FF1/FF3 from the NIST SP800-38G standard) which improve upon previous work in terms of amortized complexity in multi-target scenarios, where multiple ciphertexts are to be decrypted. Our attacks are also qualitatively better in that they make no assumptions on the correlation between the targets to be decrypted and the known plaintexts. We also surface a new vulnerability specific to FF3 and how it handles odd length domains, which leads to a substantial speedup in our attacks.

We also show the first attacks against non-Feistel based FPEs. Specifically, we show a strong message-recovery attack for FNR, a construction proposed by Cisco which replaces two rounds in the Feistel construction with a pairwise-independent permutation, following the paradigm by Naor and Reingold (JoC, '99). We also provide a strong ciphertext-only attack against a variant of the DTP construction by Brightwell and Smith, which is deployed by Protegrity within commercial applications.

All of our attacks show that existing constructions fall short of achieving desirable security levels. For Feistel and the FNR schemes, our attacks become feasible on small domains, e.g., 8 bits, for suggested round numbers. Our attack against the DTP construction is practical even for large domains. We provide proof-of-concept implementations of our attacks that verify our theoretical findings.

Keywords: Format-preserving encryption, attacks

1 Introduction

A *format-preserving encryption* (FPE) scheme is a deterministic symmetric encryption mechanism which preserves the format of the data, i.e., the ciphertext has the same format as the plaintext. For instance, a valid SSN is encrypted into a valid SSN, a valid credit-card number is encrypted into a valid credit-card

number, etc. The first known constructions date back to Brightwell and Smith [6] and Black and Rogaway [4], and a formal treatment was later given by Bellare, Ristenpart, Rogaway, and Stegers [2]. The widespread interest in FPE from industry stems for its usage in the financial sector to encrypt credit-card numbers, as well as its ability to add encryption to legacy databases and applications without violating existing format constraints. FPE has been used and deployed by several companies, e.g., Voltage, Veriphone, Ingenico, Protegrity, Cisco, as well as by major credit-card payment organizations. While precise numbers are not known, it is safe to assume that vast amounts of data are currently encrypted with FPE in industrial settings.

However, building secure FPE is a challenging question, largely because (1) the domain is usually non-binary, and standard cryptographic primitives, e.g., AES, operate on fixed-length binary domains, and (2) the domain can be *small*, and it is hard to devise schemes where the domain size is not a security parameter. For example, the ANSI ASC X9.124 standard adopted by the financial industry envisions applications with domains as small as two decimal digits. While provably-secure schemes *do* exist [11, 16, 13], they consistently fail to meet practical efficiency demands. Consequently, practical designs have been validated via cryptanalysis only, and NIST has recently standardized [9] two constructions, FF1 [3] and FF3 [5], both based on Feistel networks. Recent works have however cast some doubt on the security of these constructions, which appear to be far from the initial desiderata set by NIST's selection process, which required 128 bits of security. (Indeed, one construction, FF2 [17], was dropped for far less severe attacks [10] than those by now known to exist against all Feistel-based constructions.) This state of affairs is particularly alarming, given the large-scale usage of FPE.

In a nutshell, this paper will take FPE cryptanalysis even further, providing more evidence that practical FPE constructions with high security are still beyond reach. This is particularly important as existing standards (NIST SP 800-38G, ANSI ASC X9.124) are being revised in view of recent attacks. We will strengthen prior attacks, and also present new attacks against practical constructions (employed in industry) which do not follow the Feistel paradigm.

Existing cryptanalysis. Let us first review recent cryptanalytic attacks against FPE. Formally, an FPE scheme F is a pair of deterministic algorithms $(\mathsf{F.E}, \mathsf{F.D})$, where $\mathsf{F.E} : \mathsf{F.Keys} \times \mathsf{F.Twk} \times \mathsf{F.Dom} \to \mathsf{F.Dom}$ is the encryption algorithm, $\mathsf{F.D} : \mathsf{F.Keys} \times \mathsf{F.Twk} \times \mathsf{F.Dom} \to \mathsf{F.Dom}$ the decryption algorithm, $\mathsf{F.Keys}$ the key space, $\mathsf{F.Twk}$ the tweak space, and $\mathsf{F.Dom}$ the domain. For every key $K \in \mathsf{F.Keys}$ and tweak $T \in \mathsf{F.Twk}$, the map $\mathsf{F.E}(K, T, \cdot)$ is a permutation over $\mathsf{F.Dom}$, and $\mathsf{F.D}(K, T, \cdot)$ reverses $\mathsf{F.E}(K, T, \cdot)$.

Bellare, Hoang, and Tessaro (BHT) [1] recently introduced a framework for known-plaintext message-recovery attacks on FPE. More concretely, they introduce the notion of a *message sampler*, an algorithm XS that returns a tuple $((T_1, X_1), \ldots, (T_Q, X_Q), Z^*, a)$ that consists of Q *distinct* tweak-message pairs (T_i, X_i), a *target message* Z^*, and (possibly) some *auxiliary information*

	Our attack LD	**BHT's attack** [1]
Advantage	$1 - 1/N$	$1 - 2/N$
Running time	$O(n^2N^{r-2} + N^{r-2}np)$	$O(n \cdot N^{r-2})$
Total ciphertexts	$O(n^2N^{r-2} + N^{r-3}np)$	$O(n \cdot N^{r-2})$
Time per target	$O(n \cdot N^{r-2})$	$O(n \cdot N^{r-2})$
Ciphertexts per target	$O(n \cdot N^{r-3})$	$O(n \cdot N^{r-2})$
Ciphertexts per tweak	$O(n \cdot N)$	3
Known msg vs target	No correlation	Same right half

Table 1. Attack parameters and effectiveness. This is for balanced-Feistel FPE with domain $\{0,1\}^{2n}$ ($n \geq 3$) and r rounds, with $N = 2^n$. Our attack LD does not limit the number of targets p, and thus p can be $O(N^2)$. In contrast, BHT's attack can only handle a single target. Both attacks achieve high advantage, as shown in the second row. The third and fourth rows respectively show the running time and the number of ciphertexts for the attacks, with a generic number p of targets for LD, and a single target for BHT's attack. The fifth and sixth row shows the amortized time and the number of ciphertexts per target, if $p = \Omega(N^2)$. The seventh row shows the maximum number of ciphertexts per tweak that each attack requires, and the last row shows the needed correlation between known messages and the target messages for each attack.

$a \in \{0,1\}^*$. Then, an attacker against XS attempts to recover Z^* given

$$(T_1, \mathsf{F.E}(K, T_1, X_1)), \ldots (T_Q, \mathsf{F.E}(K, T_Q, X_Q)), a$$

for a secret key K. The attacker's advantage is obtained by subtracting from its success probability that of the best possible trivial attacker that only gets $T_1, \ldots, T_Q$ and a. Therefore, *any* message sampler with a corresponding attacker achieving substantial advantage within feasible computational constraints is effectively a break, since the scheme fails to satisfy some ideal property to be expected.

For example, for the *balanced* r-round Feistel construction with domain $\mathbb{Z}_N \times \mathbb{Z}_N$ (meaning the domain size is N^2), where $N = 2^n$, BHT provide a sampler and an attack which succeed with $O(n \cdot N^{r-2})$ ciphertexts, where in particular these ciphertexts consist of the encryption of three messages (one of which is the target one) under $O(n \cdot N^{r-2})$ distinct tweaks.[4] (The attack is summarized in Table 1.) While the attack is generic, when applied to the setting of NIST's standardized constructions FF1/FF3, which use $r = 10$ and $r = 8$, respectively, the attack becomes particularly threatening for small domains. The fact that the number of ciphertexts is larger than the domain size N is no contradiction – the point is that the number of ciphertexts *per tweak* is small, and this makes a generic message recovery without the ciphertexts only possible with small probability.

[4] BHT actually give three attacks with different complexity, but only one of them can fully recover the target message; the other two can only recover a half of the target. Since our attack can recover all target messages in their entirety, here we only compare our attack with the Full-Message Recovery attack of BHT.

We also point out the work by Durak and Vaudenay (DV) [8]. They give a message-recovery attack against FF3 which uses only *two* tweaks, yet their attack is due to a flaw in the tweaking mechanism used in FF3, rather than being a generic issue of Feistel. In contrast, BHT's attacks succeed even if the flaw behind DV's attack is fixed.

NIST has temporarily discouraged the use of FF3 as the result of DV's attack[5], whereas a draft update of the ANSI ASC X9.124 standard additionally suggests double encryption on small domains as a result of BHT's attacks.

Our contributions. The BHT attacks can be mitigated by increasing the number of rounds of the constructions. However, this raises the question of whether the attacks are the best possible, and whether new, stronger attacks, are possible. Similarly, plain Feistel is not the only approach used in practice for FPE. For example, Cisco presented a variant of Feistel, called FNR [7], which appears to bypass the BHT attacks. Protegrity is another very active company in the FPE domain and uses a different construction [12], called DTP (from "Data-type preserving" encryption), based on Brightwell and Smith's [6] construction. It is well possible that these constructions are not affected by attacks, and may end up being superior to NIST-standardized constructions.

Our first contribution will be new attacks against Feistel-based FPE that improve upon BHT in settings where multiple messages can be recovered, as well as only requiring weaker correlations in the known messages for which the FPE construction is evaluated. We will then provide an attack against FNR, thus showing it too fails to provide sufficient security. Finally, we provide a strong ciphertext-only attack against DTP. In particular, while our attacks against Feistel and FNR relies on weaknesses for small domains, our attack against DTP works even on large domains.

We complement our attacks with proof-of-concept implementations that validate experimentally our theoretical findings.

New attacks against Feistel-based FPE. We strengthen the attacks from BHT by considering the setting where the attacker is given *multiple* target messages $Z_1^*, \ldots, Z_p^*$ it is trying to recover. This captures for example an attempt by the attacker to compromise a large fraction of an FPE-encrypted database, as opposed to an individual record in it. Clearly, this task should be harder than recovering a single target, and a good FPE scheme should guarantee that the cost of recovering p messages is roughly p times that of recovering one message. Indeed, this is true when mounting BHT's attacks, as the only option is to apply the attack to each target.

We will show however that for the r-round Feistel construction with domain $\mathbb{Z}_M \times \mathbb{Z}_N$, multiple targets can be recovered much faster, in fact with a number of ciphertexts comparable to what is needed for a single target. As summarized in Table 1, for the special case $M = N = 2^n$, the amortized number of ciphertexts *per target* is only $O(n \cdot N^{r-3})$, as opposed to $O(n \cdot N^{r-2})$ when using BHT repeatedly. A further advantage of our attack is that the known plaintexts revealed

[5] `https://csrc.nist.gov/News/2017/Recent-Cryptanalysis-of-FF3`

to the attacker are not correlated with the target messages – whereas BHT assumed a fairly artificial setting where (partially) known plaintexts exhibit strong correlations with the target message.

More concretely, the attacker is supplied τ *known* distinct messages $X_1, \ldots, X_\tau$, and we have p targets $Z_1, \ldots, Z_p$. Then, the attacker gets encryptions of these $\tau + p$ messages (assumed to be distinct) under q known tweaks $T_1, \ldots, T_q$ (thus, the attacker sees $q \times (\tau + p)$ ciphertexts). The goal is to recover all of $Z_1, \ldots, Z_p$. The only assumptions here are that (1) The right halves of $X_1, \ldots, X_\tau$ cover all of $\mathbb{Z}_N$, and (2) $Z_1, \ldots, Z_p$ have (as a tuple) sufficient min-entropy conditioned on $X_1, \ldots, X_\tau, T_1, \ldots, T_q$, say at least θ. Because of this, the probability that an ideal adversary that does not learn the ciphertexts recovers all of $Z_1, \ldots, Z_p$ here is at most $2^{-\theta}$. In contrast, we give an attack which recovers them with high probability whenever q is large enough. See Table 1 for the exact complexities when $M = N = 2^n$.

We stress that unlike the BHT attacks, the attacker is not aware of any correlation between the known plaintexts $X_1, \ldots, X_\tau$ and the target plaintexts $Z_1, \ldots, Z_p$. Of course, every right half of $Z_1, \ldots, Z_p$ will appear among $X_1, \ldots, X_\tau$, but the attacker does not know which of the inputs have matching right halves. Also, we point out that the restriction of all right halves appearing in $X_1, \ldots, X_\tau$ is not as artificial as it may at first appear. If these inputs are drawn uniformly at random (under the constraint of being distinct), and $\tau = \Theta(Nn)$, then we can show that all right halves are going to appear with high probability by the so-called "coupon collector" argument. Even more importantly, if they do not cover all of $\mathbb{Z}_N$, our attacks recovers all of the $Z_1, \ldots, Z_p$ whose right halves overlap with those of $X_1, \ldots, X_\tau$.

The danger of asymmetry We note that the complexity of our attack is not symmetric in M and N. In particular, the attack's performance improves with a smaller N and a larger M. This is particularly problematic for FF3, which in the case of odd-length domains (e.g., $\{0, \ldots, 9\}^3$) would exactly create such a convenient asymmetry, setting $M = 100$ and $N = 10$. This feature was already present in the left-half attack of BHT, but went unobserved.

The FNR construction. Cisco proposed the FNR construction [7] as an approach to encrypt IP addresses. While we are not aware whether FNR was indeed used, it adopts a potentially interesting idea which seemingly prevents our and BHT's attacks against Feistel. Essentially, it uses Naor and Reingold's [15] idea of replacing the two outer rounds of the Feistel construction with a pairwise independent permutation while retaining security.

Initially, it is not clear how existing attacks against Feistel can be used when a pairwise-independent permutation is used. We show however that this approach too fails, and in fact, in terms of our attacks, FNR with r-rounds appear to be as secure as plain Feistel with $r + 2$ rounds, somehow matching (though in a different and unexplored context) the initial intuition by Naor and Reingold.

The DTP scheme and its insecurity. Another solution is the DTP scheme put forward by Protegrity [12], which is a variation of the scheme by Smith and

Brightwell [6] and which has been argued to be potentially superior to FPE.[6] In particular, reframing it in our language, DTP requires a distinct tweak per encryption, thus potentially achieving higher security by preventing detection of equal plaintexts being encrypted. However, we give an attack that only requires multiple encryptions of the same target message with different tweaks (and is thus compatible with the envisioned usage scenario). The attack differs from those against Feistel-based FPE, but again is in the same spirit of using encryptions under multiple tweaks to amplify subtle statistical deviations. We have confirmed that a variant of this scheme, called DTP-2, is still deployed by Protegrity, even though it is being phased out to be replaced with FF1.[7]

Abstractly, the main issue of DTP is that it encrypts individual digits of the plaintext $x_1x_2\dots x_n$ (where $x_i \in \mathbb{Z}_d$) as $c_i \leftarrow x_i + z_i \pmod d$, where the z_i's are pseudorandom elements of $\mathbb{Z}_D$. For example, one could use $d = 10$ (to encrypt decimal numbers) and $D = 256$ (e.g., the z_i's are individual bytes from an AES output). Then, it is not hard to see that the c_i values are not pseudorandom anymore, and there is in fact a noticeable statistical deviation. This is because $z_i \in \{0, 1, \dots, 5\}$ is more likely to occur than $z_i \in \{6, \dots, 9\}$. Our recent interactions with Protegrity indicate that $d = 62$ is more commonly used (to accommodate for the alphabet $\{a, \dots, z, A, \dots, Z, 0, \dots, 9\}$), and this introduces even more important biases. As we show below in Table 4, there is a factor 10 improvement in the number of ciphertexts required by our attack when switching from $d = 10$ to $d = 62$.

Our attack is stronger than those against Feistel and FNR as it also works on large input spaces – the problem being exploited here is the mapping between binary outputs (corresponding to the choice of D) to elements in another alphabet (by reducing mod d). The observation that encryptions are biased is not novel (cf. e.g. `https://en.wikipedia.org/wiki/Format-preserving_encryption`), but our attacks highlights how such biases can be exploited for full-message recovery in a multi-tweak scenario.

We note that the spec (as well as the original description in [6]) allow for some key-dependent pre-processing of the plaintext which Protegrity makes *explicitly optional* if tweaks are chosen uniformly at random. The version without pre-processing is the version we attack here. With pre-processing, our attack does not apply, but note that [6] acknowledges the pre-processing itself only suffices to deter "casual attacks" and this is unlikely a strong countermeasure.

ERRATA. In the proceedings version, we used a buggy variant of the coupon-collector argument. As a result, we incorrectly claimed that the attacks on Feistel-based FPE and FNR need just $\tau = \Theta(N\sqrt{n})$ known random messages. This bug is fixed in this version, by using the classical coupon-collector result [14].

[6] `http://www.protegrity.com/role-of-standards-nist-data-security/`
[7] The findings of this paper have been in particular shared with Protegrity.

2 Preliminaries

NOTATION. We let ε denote the empty string. If y is a string then $|y|$ denotes its length and $y[i]$ denotes its i-th bit for $1 \le i \le |y|$. If X is a finite set, we let $x \xleftarrow{\$} X$ denote picking an element of X uniformly at random and assigning it to x. Algorithms may be randomized unless otherwise indicated. Running time is worst case. If A is an algorithm, we let $y \leftarrow A(x_1, \ldots; r)$ denote running A with random coins r on inputs $x_1, \ldots$ and assigning the output to y. We let $y \xleftarrow{\$} A(x_1, \ldots)$ be the result of picking r at random and letting $y \leftarrow A(x_1, \ldots; r)$. We let $[A(x_1, \ldots)]$ denote the set of all possible outputs of A when invoked with inputs $x_1, \ldots$. By $\Pr[\mathrm{G}]$ we denote the probability of the event that the execution of game G results in the game returning true. If D is a set then $\mathrm{Perm}(\mathsf{D})$ denotes the set of all permutations on D. Let $\exp(x)$ denote e^x, where e is the base of the natural logarithm.

FPE. An FPE scheme F specifies a pair of deterministic algorithms $(\mathsf{F.E}, \mathsf{F.D})$, where $\mathsf{F.E} : \mathsf{F.Keys} \times \mathsf{F.Twk} \times \mathsf{F.Dom} \to \mathsf{F.Dom}$ is the encryption algorithm, $\mathsf{F.D} : \mathsf{F.Keys} \times \mathsf{F.Twk} \times \mathsf{F.Dom} \to \mathsf{F.Dom}$ the decryption algorithm, $\mathsf{F.Keys}$ the key space, $\mathsf{F.Twk}$ the tweak space, and $\mathsf{F.Dom}$ the domain. For every key $K \in \mathsf{F.Keys}$ and tweak $T \in \mathsf{T}$, the map $\mathsf{F.E}(K, T, \cdot)$ is a permutation over $\mathsf{F.Dom}$, and $\mathsf{F.D}(K, T, \cdot)$ reverses $\mathsf{F.E}(K, T, \cdot)$.

CHERNOFF BOUND. Our results heavily rely on the well-known Chernoff bounds. We recall the details of Chernoff bounds below.

Lemma 1 (Chernoff bounds). *Let $Y_1, \ldots, Y_\ell$ be independent Bernoulli random variables with $\Pr[Y_1 = 1] = \cdots = \Pr[Y_\ell = 1] = \mu$. Then,*

$$\Pr\Big[Y_1 + \cdots + Y_\ell \ge (1+\epsilon)\ell\mu\Big] \le \exp\Big(\frac{-\epsilon^2\ell\mu}{2+\epsilon}\Big) \text{ for any } \epsilon > 0, \text{ and}$$

$$\Pr\Big[Y_1 + \cdots + Y_\ell \le (1-\epsilon)\ell\mu\Big] \le \exp\Big(\frac{-\epsilon^2\ell\mu}{2}\Big) \text{ for any } 0 < \epsilon < 1.$$

3 Message recovery framework

Here we give a new formalization of message-recovery attacks, generalizing the definition of Bellare, Hoang, and Tessaro (BHT) [1] for attacking multiple target messages.

A HIGH-LEVEL INTUITION. Under our framework, there are τ known messages and p target messages. An adversary $\mathcal{A}$ will receive the ciphertexts of those, each under multiple tweaks, and has to recover at least $d \le p$ targets to win the game, where d is a parameter of the message-recovery game. For example $d = 1$ means that as long as the adversary recovers a single target message, it wins the game, and $d = p$ means that the adversary has to recover all targets to win.

Following BHT, we aim for a generalized framework that can capture BHT's attack, where known messages are correlated with the targets. Thus in our notion, the known messages and the target messages, and also the tweaks, are generated via a message sampler XS. The adversary $\mathcal{A}$ receives the tweaks and the ciphertexts, and some auxiliary information that contains information about the known messages, and possibly some partial information about the targets. We stress that only the sampler knows the target messages, and the adversary $\mathcal{A}$ just knows some partial information of the target messages that the auxiliary information reveals.

The framework above allows samplers that output target messages that are trivial to guess. Thus for any FPE scheme, there is an adversary that with high probability can recover target messages produced by those degenerate samplers by merely guessing, but of course this does not imply a vulnerability of the FPE scheme. Following BHT, we define the d-target advantage $\mathbf{Adv}^{\mathsf{mr}}_{\mathsf{F},\mathsf{XS},d}(\mathcal{A})$ of adversary $\mathcal{A}$ against FPE scheme F and sampler XS as the difference between (i) the chance that $\mathcal{A}$ can recover at least d targets, and (ii) the probability of the best strategy of guessing that many targets given just the auxiliary information (but not the ciphertexts). Hence for an FPE scheme F, if one can construct an efficient adversary $\mathcal{A}$ and an efficient sampler XS such that $\mathbf{Adv}^{\mathsf{mr}}_{\mathsf{F},\mathsf{XS},d}(\mathcal{A})$ is large, it means that this particular FPE scheme F is indeed vulnerable.

Our notion only models non-adaptive attacks and requires adversaries to recover at least d targets. However, recall that here we are giving an attack notion, and thus these restrictions only make our attacks better. On the other hand, if an FPE scheme meets our notion, it does not necessarily mean that the scheme is secure for real-world usage. Below, we will formalize our framework.

Samplers and guessing probability. A *message sampler* is an algorithm XS that returns $((T_1, X_1), \ldots, (T_Q, X_Q), Z_1, \ldots, Z_p, a)$ that consists of Q tweak-message pairs (T_i, X_i), p *target messages* Z_j, and some *auxiliary information* $a \in \{0,1\}^*$. Note that encryption schemes of FPEs are deterministic, and thus it is trivial to detect repetition among the pairs $(T_1, X_1), \ldots, (T_Q, X_Q)$ given their ciphertexts. Therefore, following BHT, we require the *distinctness* condition that the Q pairs $(T_1, X_1), \ldots, (T_Q, X_Q)$ be distinct. Define the d-target message-guessing (mg) advantage against a sampler XS as

$$\mathbf{Adv}^{\mathsf{mg}}_{\mathsf{XS},d} = \max_{\mathcal{S}} \Pr[\mathbf{G}^{\mathsf{mg}}_{\mathsf{XS},d}(\mathcal{S})],$$

where game $\mathbf{G}^{\mathsf{mg}}_{\mathsf{XS}}(\mathcal{S})$ is defined in the top panel of Fig. 1. This is the probability of the best possible way at guessing at least d target messages given the tweaks and auxiliary information. For the special case $d = p$, meaning that one has to guess all target messages, we write $\mathbf{Adv}^{\mathsf{mg}}_{\mathsf{XS}}$ instead of $\mathbf{Adv}^{\mathsf{mg}}_{\mathsf{XS},p}$. To account for the efficiency of attacks, besides the number of ciphertexts Q, we also consider the *number of ciphertexts per recovered target* $q_t = Q/d$. This is the amortized data complexity.

Message-recovery notion. Let F be an FPE scheme. Let XS be a message sampler such that $T_1, \ldots, T_Q \in \mathsf{F.Twk}$ and $X_1, \ldots, X_Q, Z_1, \ldots, Z_p \in \mathsf{F.Dom}$ for

Game $\mathbf{G}^{\mathsf{mg}}_{\mathsf{XS},d}(\mathcal{S})$

$((T_1, X_1), \ldots, (T_Q, X_Q), Z_1, \ldots, Z_p, a) \xleftarrow{\$} \mathsf{XS}$

$(Z_1^*, \ldots, Z_p^*) \xleftarrow{\$} \mathcal{S}(T_1, \ldots, T_Q, a)$; $t \leftarrow \min\{d, p\}$

Return $(\exists\, i_1 < \cdots < i_t$ such that $(Z_{i_1} = Z_{i_1}^*) \wedge \cdots \wedge (Z_{i_t} = Z_{i_t}^*))$

Game $\mathbf{G}^{\mathsf{mr}}_{\mathsf{F},\mathsf{XS},d}(\mathcal{A})$

$K \xleftarrow{\$} \mathsf{F.Keys}$; $((T_1, X_1), \ldots, (T_Q, X_Q), Z_1, \ldots, Z_p, a) \xleftarrow{\$} \mathsf{XS}$

For $i = 1, \ldots, Q$ do $Y_i \leftarrow \mathsf{F.E}(K, T_i, X_i)$

$(Z_1^*, \ldots, Z_p^*) \xleftarrow{\$} \mathcal{A}((T_1, Y_1), \ldots, (T_Q, Y_Q), a)$; $t \leftarrow \min\{d, p\}$

Return $(\exists\, i_1 < \cdots < i_t$ such that $(Z_{i_1} = Z_{i_1}^*) \wedge \cdots \wedge (Z_{i_t} = Z_{i_t}^*))$

Fig. 1. Games defining message-recovery notion of an FPE scheme F, parameterized by a message sampler XS.

any $((T_1, X_1), \ldots, (T_Q, X_Q), Z_1, \ldots, Z_p, a)$ in $[\mathsf{XS}]$. Define the d-target message-recovery (mr) advantage of $\mathcal{A}$ against F, XS as

$$\mathbf{Adv}^{\mathsf{mr}}_{\mathsf{F},\mathsf{XS},d}(\mathcal{A}) = \Pr[\mathbf{G}^{\mathsf{mr}}_{\mathsf{F},\mathsf{XS},d}(\mathcal{A})] - \mathbf{Adv}^{\mathsf{mg}}_{\mathsf{XS},d}\,.$$

The mr game $\Pr[\mathbf{G}^{\mathsf{mr}}_{\mathsf{F},\mathsf{XS},d}(\mathcal{A})]$ is defined in the bottom panel of Fig. 1, measuring $\mathcal{A}$'s advantage at recovering at least d target messages given the tweaks, ciphertexts, and auxiliary information. For $d = p$, meaning that the adversary has to recover all targets, we write $\mathbf{Adv}^{\mathsf{mr}}_{\mathsf{F},\mathsf{XS}}(\mathcal{A})$ instead of $\mathbf{Adv}^{\mathsf{mr}}_{\mathsf{F},\mathsf{XS},p}(\mathcal{A})$.

Relation to BHT's notion. BHT's notion is the special case of the definition above where the number of target message p is 1. However, in practice, it is not economical to collect a lot of known message-ciphertext pairs to recover just a single target message. If we can instead spend the same amount of resource but recover multiple messages, the cost will be amortized by the number of recovered targets, cheapening the attack. Thus compared to BHT's definition, ours gives a more realistic attack model.

Remarks. Most existing notions in the cryptanalytic literature only define codebook-recovery attacks, but our attacks or BHT's attack do not fit into this category. Bellare, Ristenpart, Rogaway, and Stegers (BRRS) [2] define a message-recovery notion for FPEs, but again (i) this notion considers just a single target message, and (ii) more importantly, the number of ciphertexts under this notion cannot exceed the domain size. Thus BRRS's notion also fails to capture our attack or BHT's attack.

4 Attacking Feistel-based FPE

In this section, we first recall the Feistel-based FPE constructions, as in NIST standards FF1 or FF3, and then give a message-recovery attack on a generic FPE scheme. Compared to BHT's attacks [1], our attack can deal with a general number of target messages and recover all of them, and thus have better

$\underline{\mathsf{F.E}(K,T,X)}$
$(L,R) \leftarrow X$
For $i = 1$ to r do
  If $(i \bmod 2 = 1)$ then $L \leftarrow L \boxplus F_i(K,T,R)$
  Else $R \leftarrow R \boxplus F_i(K,T,L)$
Return (L,R)

$\underline{\mathsf{F.D}(K,T,Y)}$
$(L,R) \leftarrow Y$
For $i = r$ to 1 do
  If $i \bmod 2 = 1$ then $L \leftarrow L \boxminus F_i(K,T,R)$
  Else $R \leftarrow R \boxminus F_i(K,T,L)$
Return (L,R)

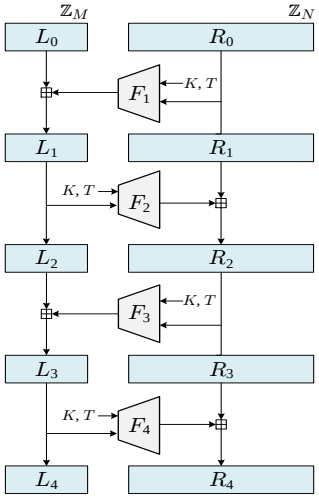

Fig. 2. Left: The code for the encryption and decryption algorithms of $\mathsf{F} = \mathbf{Feistel}[r, M, N, \boxplus, \mathrm{PL}]$, where $\mathrm{PL} = (\mathcal{T}, \mathcal{K}, F_1, \ldots, F_r)$. **Right:** An illustration of encryption with $r = 4$ rounds.

amortized cost. Moreover, we do not require any correlation between the known messages and the targets.

Feistel-based constructions. Most existing FPE schemes, including the FF1 and FF3 standards [9], are based on Feistel networks. Following BHT, we specify Feistel-based FPE in a general, parameterized way. This allows us to refer to both schemes of ideal round functions for the analysis, and schemes of some concrete round functions for realizing the standards.

We associate to parameters $r, M, N, \boxplus, \mathrm{PL}$ an FPE scheme $\mathsf{F} = \mathbf{Feistel}[r, M, N, \boxplus, \mathrm{PL}]$. Here $r \geq 2$ is an integer, the number of rounds, and $\boxplus$ is an operation for which $(\mathbb{Z}_M, \boxplus)$ and $(\mathbb{Z}_N, \boxplus)$ are Abelian groups. We let $\boxminus$ denote the inverse operator of $\boxplus$, meaning that $(X \boxplus Y) \boxminus Y = X$ for every X and Y. Integers $M, N \geq 1$ define the domain of F as $\mathsf{F.Dom} = \mathbb{Z}_M \times \mathbb{Z}_N$. The parameter $\mathrm{PL} = (\mathcal{T}, \mathcal{K}, F_1, \ldots, F_r)$ specifies the set $\mathcal{T}$ of tweaks and a set $\mathcal{K}$ of keys, meaning $\mathsf{F.Twk} = \mathcal{T}$ and $\mathsf{F.Keys} = \mathcal{K}$, and the round functions $F_1, \ldots, F_r$ such that $F_i : \mathcal{K} \times \mathcal{T} \times \mathbb{Z}_N \to \mathbb{Z}_M$ if i is odd, and $F_i : \mathcal{K} \times \mathcal{T} \times \mathbb{Z}_M \to \mathbb{Z}_N$ if i is even. The code of $\mathsf{F.E}$ and $\mathsf{F.D}$ is shown in Fig. 2.

Classical Feistel schemes correspond to the boolean case, where $M = 2^m$ and $N = 2^n$ are powers of two, and $\boxplus$ is the bitwise xor operator $\oplus$. The scheme is balanced if $M = N$ and unbalanced otherwise. For $X = (L, R) \in \mathbb{Z}_M \times \mathbb{Z}_N$, we call L and R the *left segment* and *right segment* of X, respectively. We write $\mathbf{Left}(X)$ and $\mathbf{Right}(X)$ to refer to the left and right segments of X respectively. For simplicity, we assume that 0 is the zero element of the groups $(\mathbb{Z}_M, \boxplus)$ and $(\mathbb{Z}_N, \boxplus)$.

For analysis, the round functions are modeled as truly random. Formally, let $\mathcal{T} = \{0,1\}^*$, and let $\mathcal{K}$ be the set $\mathbf{RF}(\mathcal{T}, r, M, N)$ of all tuples of functions $(G_1, \ldots, G_r)$ such that $G_i : \mathcal{T} \times \mathbb{Z}_N \to \mathbb{Z}_M$ if i is odd, and $G_i : \mathcal{T} \times \mathbb{Z}_M \to \mathbb{Z}_N$ if i is even. Then for $1 \le i \le r$ define $F_i(K, \cdot, \cdot) = G_i(\cdot, \cdot)$, where $(G_1, \ldots, G_r) \leftarrow K$. We write $\mathbf{Feistel}[r, M, N, \boxplus]$ to denote $\mathbf{Feistel}[r, M, N, \boxplus, \mathrm{PL}]$, for the particular choice $\mathrm{PL} = (\mathcal{T}, \mathcal{K}, F_1, \ldots, F_r)$ above.

Schemes in the standards [9] specify the round functions using AES. Using the standard assumption that AES is a PRF, one can focus on attacking Feistel-based schemes of ideal round functions, with small differences in the advantage.

Setup. We give a message-recovery attack on a generic Feistel-based FPE $\mathsf{F} = \mathbf{Feistel}[r, M, N, \boxplus, \mathrm{PL}]$. Like the prior work of BHT [1], we only consider the case that r is even, as NIST standards only use $r = 8$ (for FF3) or $r = 10$ (for FF1). Under our attack, there are τ known messages $X_1, \ldots, X_\tau$ and p targets $Z_1, \ldots, Z_p$. The adversary is given the encryption of those $\tau + p$ distinct messages under q tweaks $T_1, \ldots, T_q$, for an appropriately large q. Due to the distinctness requirement, $X_1, \ldots, X_\tau, Z_1, \ldots, Z_p$ must be distinct. The auxiliary information is $(X_1, \ldots, X_\tau, p, q)$. The only requirement in our attack is that with high probability, the right halves of the known messages $X_1, \ldots, X_\tau$ cover at least d of the right halves of the targets. We have no restriction on the number p of targets or the parameter d, (except the unavoidable constraint that $d \le p$) so potentially p can be as large as $MN - \tau$. Our attack will recover d targets out of $Z_1, \ldots, Z_p$.

A special important case in our attack is that the right halves of $X_1, \ldots, X_\tau$ cover everything in $\mathbb{Z}_N$; in this case we can recover all targets. At the first glance, this requirement seems contrived, and thus it is unclear how the adversary can mount such an attack. However, we will show that for $\tau = \lceil N(\ln(N)+3)\} \rceil$, if the known messages are sampled uniformly without replacement from $\mathbb{Z}_M \times \mathbb{Z}_N$ then they will meet the requirement above with probability at least 0.95. Concretely, if we want to recover PINs, meaning $M = N = 100$, we need to obtain 761 random known messages. In contrast, BHT's attack needs to obtain two known messages, but one of those must have the same right half as the target.

To explain the bound $\lceil \min\{N(\ln(N)+3)\} \rceil$ above, note that this is the well-known coupon collector's problem: there are N types of coupons and a collector wishes to collect all of them. In the classical setting, for each draw the collector obtains a uniformly random type. In contrast, in our settings, because $Z_1, \ldots, Z_p$ are distinct, each time the collector buys a coupon, its type is slightly biased towards new types that the collector has not yet owned. This means that while the classical bound, stated in Lemma 2 below, continues to apply to our setting, we might need fewer coupons than what is suggested in the classical setting.

Lemma 2 (Coupon collector's problem). *[14, Chapter 3.6] Let $N \ge 1$ be an integer, and let $\lambda > 0$ be a real number. Suppose that there are N types of coupon and a collector buys $\tau = \lceil N(\ln(N) + \lambda) \rceil$ coupons of truly random types. Then the chance that the collector gets all N types is at least $1 - e^{-\lambda}$.*

From Lemma 2 above, the requirement of our attack is quite mild, yet it is powerful, recovering as many targets as possible. In contrast, in BHT's attack, there is only a single target (meaning $p = 1$), and the first known message must have the same right half as the target message. Of course in our attack, for each target Z_i, there is some known message X_j of the same right half as Z_i, but the adversary does not know what is j.

THE ATTACK. We formalize the attack via the message-recovery framework, by specifying a class $\mathsf{SC1}_{p,q,\delta,\theta}$ of samplers, and then giving a lower bound on the mr-advantage of the attack for any sampler in this class. First, let $\mathsf{DC1}_{p,q,d,\delta,\theta}$ be the class of all algorithms D that outputs q distinct tweaks $T_1, \ldots, T_q \in \{0,1\}^*$, and distinct $X_1, \ldots, X_\tau, Z_1, \ldots, Z_p \in Z_M \times \mathbb{Z}_N$ such that (1) with probability at least $1-\delta$, there are d or more indices k such that $Z_k \in \{\mathbf{Right}(X_1), \ldots, \mathbf{Right}(X_\tau)\}$ and (2) given $X_1, \ldots, X_\tau, T_1, \ldots, T_q$, for any subset $\{r_1, \ldots, r_d\} \subseteq \{1, \ldots, \tau\}$, for any $Z_1^*, \ldots, Z_d^* \in \mathbb{Z}_M \times \mathbb{Z}_N \backslash \{X_1, \ldots, X_\tau\}$, the conditional probability that $Z_{r_1} = Z_1^*, \ldots, Z_{r_d} = Z_d^*$ is at most $2^{-\theta}$.[8] To any such D, we associate the sampler

Sampler $\mathsf{XS}[\mathsf{D}]$
$(T_1, \ldots, T_q, X_1, \ldots, X_\tau, Z_1, \ldots, Z_p) \leftarrow\!\!\$\ \mathsf{D}$
$a \leftarrow (X_1, \ldots, X_\tau, p, q)$
Return $\big(\{(T_i, X_j), (T_i, Z_k) \mid i \le q, j \le \tau, k \le p\}, Z_1, \ldots, Z_p, a\big)$

The sampler above returns the pairs (T_i, X_j) and (T_i, Z_k) for every $i \le q$ and every $j \le \tau$, and $k \le p$, where the targets are $Z_1, \ldots, Z_p$. The number of ciphertexts Q is $(\tau + p)q$, and the number of ciphertexts per recovered target q_t is $(\tau+p)q/d$. Let $\mathsf{SC1}_{p,q,d,\delta,\theta} = \{\mathsf{XS}[\mathsf{D}] \mid \mathsf{D} \in \mathsf{DC1}_{p,q,d,\delta,\theta}\}$. We would expect that adversaries will have low mr-advantage, even if q is big. However, the Left-half Differential (LD) attack, given in Fig. 3, can recover d targets out of $Z_1, \ldots, Z_p$ in $O(pqN)$ time. Theorem 3 below gives a lower bound on the mr-advantage of LD.

The bound in Theorem 3, for the special case $d = p$, is illustrated in Fig. 4. For example, for FF1, the attack is only reasonably feasible in very few domains, say one-byte strings ($M = N = 16$) or two-decimal strings ($M = N = 10$), but recall that FF1 and FF3 are supposed to provide 128-bit security whenever the domain size MN is at least 100. For FF3, since there are fewer rounds, the attack is faster, and thus becomes feasible in more domains.

Theorem 3. *Let $M, N \ge 4$ and let $p, q \ge 1$ be integer. Let $r \ge 4$ be an even integer such that $N^{(r-2)/2} \ge 2M$, and let d be an integer such that $1 \le d \le p$. Let $\mathsf{F} = \mathbf{Feistel}[r, M, N, \boxplus]$, and let $\lambda = \left(1 - \frac{1}{M-1}\right)^2 \left(1 - \frac{1}{MN}\right)$. Then for any $0 \le \delta \le 1$ and any $\theta \ge 0$, and for any sampler XS in the class $\mathsf{SC1}_{p,q,d,\delta,\theta}$,*

$$\mathbf{Adv}^{\mathsf{mr}}_{\mathsf{F},\mathsf{XS},d}(\mathsf{LD}) \ge 1 - \delta - d \cdot \exp\Big(\frac{-\lambda M q}{12 \cdot N^{r-2}}\Big) - MNd \cdot \exp\Big(\frac{-\lambda M q}{9 \cdot N^{r-2}}\Big) - 2^{-\theta} \ .$$

[8] For the special case where $Z_1, \ldots, Z_p$ are sampled uniformly without replacement from $(\mathbb{Z}_M \times \mathbb{Z}_N) \backslash \{X_1, \ldots, X_\tau\}$, then $\theta = \Theta(d \cdot \log(MN))$.

Adversary $\mathsf{LD}(\{(T_i, C_{i,j}), (T_i, C'_{i,k})\}_{i,j,k}, a)$
// $1 \le i \le q, 1 \le j \le \tau, 1 \le k \le p$
$(X_1, \ldots, X_\tau, p, q) \leftarrow a$; $S, \mathrm{Dom} \leftarrow \emptyset$
For $j = 1, \ldots, \tau$ do
  If $\mathbf{Right}(X_j) \notin \mathrm{Dom}$ then $S \leftarrow S \cup \{j\}$; $\mathrm{Dom} \leftarrow \mathrm{Dom} \cup \{\mathbf{Right}(X_j)\}$
For $k \leftarrow 1$ to p do // Recover target Z_k
  For $j \in S$, $s \in \mathbb{Z}_M$ do $V_{j,s} \leftarrow 0$
  For $i \leftarrow 1$ to q, $j \in S$ do
    $s \leftarrow \mathbf{Left}(C'_{i,k}) \boxminus \mathbf{Left}(C_{i,j}) \boxplus \mathbf{Left}(X_j)$; $V_{j,s} \leftarrow V_{j,s} + 1$
  Let $V_{j^*,s^*} = \max\{V_{j,s} \mid j \in S, s \in \mathbb{Z}_M\}$; $Z_k \leftarrow (s^*, \mathbf{Right}(X_{j^*}))$
Return $(Z_1, \ldots, Z_p)$

Fig. 3. The Left-half Differential attack.

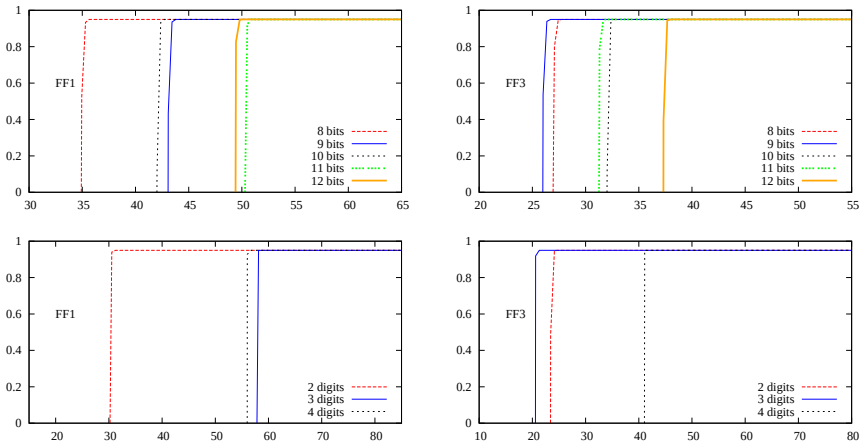

Fig. 4. The mr advantage of the Left-half Differential attack for binary strings of 8–12 bits (top) and decimal strings of 2–4 digits (bottom). The x-axis shows the log, base 2, of the number q of tweaks (which is also roughly q_t, the number of ciphertexts per recovered target), and the y-axis shows $\mathbf{Adv}^{\mathsf{mr}}_{\mathbf{Feistel}[r,M,N,\boxplus],\mathsf{XS}}(\mathsf{LD})$, for XS that outputs $\tau = \lceil \min\{N(\ln(N) + 3)\} \rceil$ known messages $X_1, \ldots, X_\tau$ and $p = MN - \tau$ targets; those MN messages are sampled uniformly without replacement from $\mathbb{Z}_M \times \mathbb{Z}_N$. Here we aim to recover all targets, namely $d = p$. On the left, we use the parameters of the FF1 standard. On the right, we use parameters of FF3.

IDEAS OF THE ATTACK. Our attack is based on an observation by BHT that for any two messages X and X' of the same right half, if we encrypt them under the same tweak to obtain ciphertexts C and C' respectively, then $\mathbf{Left}(C) \boxminus \mathbf{Left}(C')$

is most likely to be $\mathbf{Left}(X) \boxminus \mathbf{Left}(X')$. This observation is formally stated in Lemma 4 below.

Lemma 4 ([1]). *Let* $\mathsf{F} = \mathbf{Feistel}[r, M, N, \boxplus]$. *Fix distinct* $X, X' \in \mathbb{Z}_M \times Z_N$ *of the same right segment, a tweak* $T \in \mathsf{F.Twk}$, *and an even integer* $t \in \{2, 4, \ldots, r\}$. *Pick* $K \leftarrow\!\!\$\ \mathsf{F.Keys}$. *Let* L_t *and* L'_t *be the the left segment of the round-*t *output of* X *and* X' *under* $\mathsf{F}(K, T, \cdot)$, *respectively. Then*

(a) $\Pr[L_t \boxminus L'_t = L_0 \boxminus L'_0] \geq \frac{N}{MN-1} + \frac{1-1/(M-1)}{N^{(t-2)/2}}$.
(b) $\Pr[L_t \boxminus L'_t = Z] \leq \frac{N}{MN-1}$, *for any* $Z \in \mathbb{Z}_M \backslash \{L_0 \boxminus L'_0\}$.

The probabilities above are taken over a sampling $K \leftarrow\!\!\$\ \mathsf{F.Keys}$.

Consider a target Z_k such that $\mathbf{Right}(Z_k) \in \{\mathbf{Right}(X_1), \ldots, \mathbf{Right}(X_\tau)\}$.[9] Among the known messages $X_1, \ldots, X_\tau$, there will be some X_{j^*} of the same right segment as Z_k. Suppose that somehow we know j^*. Then obviously we can recover the right segment of Z_k. To recover the left segment of Z_k, we will use the above observation of BHT. For all ciphertexts C and C' of X_{j^*} and Z_k under the same tweak respectively, one can guess $\mathbf{Left}(Z_k)$ as $\mathbf{Left}(C') \boxminus \mathbf{Left}(C) \boxplus \mathbf{Left}(X_{j^*})$. However, compared to a random guessing, this is only slightly better; the improvement in the advantage is about $\frac{1-1/(M-1)}{N^{(r-2)/2}}$. To amplify the advantage, we consider ciphertexts C_i and C'_i of X_{j^*} and Z_k under many tweaks T_i, and output the majority value of those $\mathbf{Left}(C'_i) \boxminus \mathbf{Left}(C_i) \boxplus \mathbf{Left}(X_{j^*})$.

Since the algorithm above *assumes* that we are given the index j^*, we are left with the task of finding j^*. We first narrow down our search by considering a smallest possible subset S of $\{1, \ldots, \tau\}$ such that $\{\mathbf{Right}(X_j) \mid j \in S\} = \{\mathbf{Right}(X_1), \ldots, \mathbf{Right}(X_\tau)\}$. Such a set S will contain j^*, but we still do not know which is the right one, among $|S|$ possible values. Next, we try the strategy above for *every* $j \in S$ to see which gives us the best majority value. Specifically, for every $j \in S$, we consider ciphertexts $C_{i,j}$ and $C'_{i,k}$ of X_j and Z_k under tweaks T_i respectively. For every $i \in \{1, \ldots, q\}$, let $U_{i,j} \leftarrow \mathbf{Left}(C'_i) \boxminus \mathbf{Left}(C_i) \boxplus \mathbf{Left}(X_j)$. We then find the majority value of $U_{1,j}, \ldots, U_{q,j}$ together with the number V_j of its occurrences among those q values. Finally, in the election for j^*, each candidate j has V_j votes. The winner is the candidate of the most votes.

The code in Fig. 3 implements the algorithm above as follows. For each $s \in \mathbb{Z}_N$ and each $j \in S$, we count the number $V_{j,s}$ of the occurrences of s in $U_{1,j}, \ldots, U_{q,j}$. We then find (j^*, s^*) such that $V_{j^*,s^*} = \max\{V_{j,s} \mid j \in S, s \in \mathbb{Z}_N\}$. The value s^* is the left segment of Z_k, and the right segment of X_{j^*} is also the right segment of Z_k.

To justify the way we pick j^* above, we need to understand the distribution of $V_{j,s}$, for every $j \in \mathbb{Z}_N \backslash \{j^*\}$ and $s \in \mathbb{Z}_N$. Each such message X_j will have a different right segment from Z_k. The following Lemma 5 tells us that if we encrypt X_j and Z_k under the same tweak to get ciphertexts C and C' respectively,

[9] We stress that the adversary does not need to know that $\mathbf{Right}(Z_k) \in \{\mathbf{Right}(X_1), \ldots, \mathbf{Right}(X_\tau)\}$; it will blindly use the same algorithm for all targets, but will happen to recover Z_k correctly.

Domain	Our cost q_t (for FF1)	Our cost q_t (for FF3)	BHT's cost q_t (for FF1)	BHT's cost q_t (for FF3)
$\{0,1\}^8$	2^{36}	2^{27}	2^{38}	2^{30}
$\{0,1\}^9$	2^{44}	2^{26}	2^{46}	2^{38}
$\{0,\ldots,9\}^2$	2^{31}	2^{24}	2^{34}	2^{27}
$\{0,\ldots,9\}^3$	2^{58}	2^{21}	2^{62}	2^{49}

Table 2. Comparison of our Left-half Differential attack, and BHT's attack on Feistel$[r, M, N, \boxplus]$ on parameters of FF1 and FF3. The first column shows the domain $\mathbb{Z}_M \times \mathbb{Z}_N$. The second and third columns show estimated values of q_t—the number of ciphertexts per recovered target—needed for our attack, for FF1 and FF3, respectively, to achieve advantage 0.9. (For our attack, q_t is also approximately q, the number of tweaks.) We use $\tau = \lceil N(\ln(N)+3)\rceil$ known messages $X_1, \ldots, X_\tau$ and $p = MN - \tau$ targets; those MN messages are sampled uniformly without replacement from $\mathbb{Z}_M \times \mathbb{Z}_N$. Our attack aims to recover all targets, namely $d = p$. The fourth and fifth columns show estimated values of q_t needed for BHT's attack, for FF1 and FF3, respectively, to achieve advantage 0.9.

---

then $\mathbf{Left}(C') \boxminus \mathbf{Left}(C)$ is uniformly distributed over $\mathbb{Z}_M$. The proof is given in Appendix A.

Lemma 5. *Let* $\mathsf{F} = \mathbf{Feistel}[r, M, N, \boxplus]$. *Fix distinct* $X, X' \in \mathbb{Z}_M \times Z_N$ *of different right segments, a tweak* $T \in \mathsf{F.Twk}$, *and an even integer* $t \in \{2, 4, \ldots, r\}$. *Pick* $K \xleftarrow{\$} \mathsf{F.Keys}$. *Let* L_t *and* L'_t *be the the left segment of the round-*t *output of* X *and* X' *under* $\mathsf{F}(K, T, \cdot)$, *respectively. Then for any* $Z \in \mathbb{Z}_M$, *we have* $\Pr[L_t \boxminus L'_t = Z] = \frac{1}{M}$, *where the probability is taken over a random sampling* $K \xleftarrow{\$} \mathsf{F.Keys}$.

On the one hand, from Lemma 4, the expected value of V_{j^*,s^*} is at least $q(\mu + \Delta)$, where $\mu = \frac{N}{MN-1}$ and $\Delta = \frac{1-1/(M-1)}{N^{(t-2)/2}}$. On the other hand, by using Lemma 5, the expected value of each other $V_{j,s}$ is at most $q\mu$. We will show that it is unlikely for V_{j^*,s^*} to get below the threshold $q(\mu + \Delta/2)$, and any other $V_{j,s}$ is unlikely to get beyond that threshold.

Discussion. A concrete comparison of our attack and BHT's attack is shown in Table 2. When the domain length is odd, FF1 and FF3 have different ways to interpret what are M and N. For example, for domain $\{0,\ldots,9\}^3$ (namely 3-digit numbers), FF1 uses $M = 10$ and $N = 100$, whereas FF3 uses $M = 100$ and $N = 10$. An interesting observation is that in those odd domains, our attack does not improve BHT's attack for FF1, but significantly improves BHT's attack for FF3. For example, for domain $\{0,\ldots,9\}^3$ above, our attack uses $q_t = 2^{58}$ for FF1 and BHT's uses $q_t = 2^{62}$, but for FF3, our attack only needs $q_t = 2^{21}$, whereas BHT's attack requires $q_t = 2^{49}$. Thus our attack (i) shows that FF3's way of partitioning odd domains is inferior to that of FF1, and (ii) underscores that for tiny domains, the round counts that FF1 and FF3 use are not enough, as BHT's attack already pointed out. In other words, our

Domain	Number of tweaks, q	Recovery rate	Time (min)	Number of tweaks, q	Recovery rate	Time (min)
$\{0,1\}^7$	2^{20}	100%	0.9	2^{19}	66%	0.46
$\{0,\ldots,9\}^2$	2^{23}	100%	5.92	2^{22}	86%	3.06
$\{0,\ldots,9\}^3$	2^{20}	100%	8.72	2^{19}	66%	5.3

Table 3. Empirical results of our Left-half Differential attack against FF3. For each domain (shown in the first column), we run experiments with two values of q (the number of tweaks) as indicated in the second and fifth columns. The recovery rates corresponding to these two values of q are given in the third and sixth columns, respectively. Finally, the average running time (in minutes) of each experiment is given in the fourth and seventh columns.

attack surfaces weaknesses which might have eliminated these algorithms from consideration during standardization,[10] and they significantly reduce confidence in these algorithms, which are widely deployed.

The recent FF3 attack by Durak and Vaudenay (DV) [8] can recover the entire codebook for quite bigger domains, such as PINs ($M = N = 100$). However, this attack is adaptive, meaning that the adversary must choose the next known message based on prior ciphertexts, which is very hard to mount in practice. Moreover, DV's attack can be easily fixed without performance penalty by restricting the tweak space. In contrast, to thwart our attack or BHT's attack, for tiny domains one has to add a few more rounds, which is widely perceived as a drawback for performance-hungry applications.

Experiments. As a proof of concept, we implement our Left-half Differential attack, and evaluate its message-recovery rate against FF3. Each experiment was run using 64 threads in a server of Intel(R) Xeon(R) CPU E5-2699 v3 2.30GHz CPU and 256 GB RAM. Our implementation, written in Go, uses FF3 source code from Capital One.[11] We evaluate our attack on three domains: $\{0,1\}^7$ (namely $M = 16$ and $N = 8$), $\{0,\ldots,9\}^2$ (namely $M = N = 10$), and $\{0,\ldots,9\}^3$ (namely $M = 100$ and $N = 10$); each on several values of q, the number of tweaks. For each domain $\mathbb{Z}_M \times \mathbb{Z}_N$ and each choice of q, we fix τ known messages whose right segments cover $\mathbb{Z}_N$, and run the attack for 100 trials, where $\tau = 33$ for $\{0,1\}^7$, $\tau = 31$ for $\{0,\ldots,9\}^2$, and $\tau = 96$ for $\{0,\ldots,9\}^3$. While the known messages are fixed for all 100 trials, we use $p = MN - \tau$ target messages, and randomly shuffle the targets for each trial. Here we aim to recover all targets, namely $d = p$.

The results of our experiments, given in Table 3, are consistent with (and even slightly better than) Theorem 3. For example, for domain $\{0,\ldots,9\}^2$, theoretically, one would need to use about $q = 2^{24}$ tweaks to recover all targets with probability nearly 1, and our experiments confirm that using $q = 2^{24}$ indeed gives

[10] Recall that FF2 was eliminated due to a theoretical attack using 2^{64} ciphertexts.
[11] Capital One's code is available at `https://github.com/capitalone/fpe`.

100% recovery rate. However, even for $q = 2^{23}$, in every trial we can recover all targets, and the average running time to recover target messages for each trial is about 5.92 minutes. If one instead uses $q = 2^{22}$, then the recovery rate drops to 86%, meaning that in 86 out of 100 trials, we can recover all targets.

Our experiments above empirically confirm the correctness of our attack for tiny domains. Below, we will give a formal proof to rigorously justify our attack for all domains.

Proof of Theorem 3. First we show that $\mathbf{Adv}^{\mathsf{mg}}_{\mathsf{XS}} \leq 2^{-\theta}$. Consider an arbitrary simulator $\mathcal{S}$. To win the game, $\mathcal{S}$ must find the first target Z_1. The simulator is only given the tweaks and the auxiliary information $(X_1, \ldots, X_\tau, p, q)$, and has to guess correctly at least d components of $(Z_1, \ldots, Z_p)$. From the definition of θ, the chance that the simulator's guess is correct is at most $2^{-\theta}$. Next, we show that

$$\Pr[\mathbf{G}^{\mathsf{mr}}_{\mathsf{F},\mathsf{XS}}(\mathsf{LD})] \geq 1 - \delta - d \cdot \exp\Big(\frac{-\lambda Mq}{12 \cdot N^{r-2}}\Big) - MNd \cdot \exp\Big(\frac{-\lambda Mq}{9 \cdot N^{r-2}}\Big) \ .$$

Let $S \subseteq \{1, \ldots, \tau\}$ be a set such that $\{\mathbf{Right}(X_j) \mid j \in S\} = \{\mathbf{Right}(X_1), \ldots, \mathbf{Right}(X_\tau)\}$. With probability at least $1 - \delta$, at least d targets will have their right halves in $\{\mathbf{Right}(X_j) \mid j \in S\}$. Fix a target Z_k such that $\mathbf{Right}(Z_k) \in \{\mathbf{Right}(X_j) \mid j \in S\}$. By union bound, it suffices to show that the chance the adversary fails to recover Z_k is at most

$$\exp\Big(\frac{-\lambda Mq}{12 \cdot N^{r-2}}\Big) + MN \cdot \exp\Big(\frac{-\lambda Mq}{9 \cdot N^{r-2}}\Big) \ .$$

Recall that for every $j \in S$ and every $s \in \mathbb{Z}_N$, we keep track of the number $V_{j,s}$ of the occurrences of s among the values $U_{1,j}, \ldots, U_{q,j}$, where $U_{i,j} \leftarrow \mathbf{Left}(C'_{i,k}) \boxminus \mathbf{Left}(C_{i,j}) \boxplus \mathbf{Left}(X_j)$. Let j^* be the element of S such that $\mathbf{Right}(X_{j^*}) = \mathbf{Right}(Z_k)$, and let $s^* \leftarrow \mathbf{Left}(Z_k)$. The adversary can recover Z_k if V_{j^*,s^*} is the maximum of $\{V_{j,s} \mid j \in S, s \in \mathbb{Z}_N\}$. Let $\mu \leftarrow \frac{N}{MN-1}$ and $\Delta \leftarrow \frac{1-1/(M-1)}{N^{(r-2)/2}}$. We will give (i) an upper bound for the probability that $V_{j,s}$, with $(j,s) \neq (j^*, s^*)$, is bigger than the threshold $q(\mu + \Delta/2)$, and (ii) an upper bound for the probability that V_{j^*,s^*} is smaller than that threshold. Both (i) and (ii) are handled using Chernoff bounds.

Proceeding into details, fix $(j,s) \neq (j^*, s^*)$. For each $i \leq q$, let Y_i be the Bernoulli random variable such that $Y_i = 1$ if and only if $U_{i,j} = s$. The random variables $Y_1, \ldots, Y_q$ are independent and identically distributed (as they are produced from a Feistel network of ideal round functions, under distinct tweaks), and $V_{j,s} = Y_1 + \cdots + Y_q$. Let $\nu = \Pr[Y_1 = 1] \leq \mu$ and $\epsilon = \frac{\Delta}{2\nu} \geq \frac{\Delta}{2\mu}$. Note that $\Delta/\mu \leq M/N^{(r-2)/2} \leq 1/2$, and $\Delta^2/\mu = \lambda M/N^{r-2}$. Then

$$\frac{\epsilon^2 \nu}{2+\epsilon} = \frac{\Delta}{4/\epsilon + 2} \geq \frac{\Delta}{8\mu/\Delta + 2} = \frac{\Delta^2/\mu}{8 + 2\Delta/\mu} \geq \frac{\lambda M}{9 \cdot N^{r-2}} \ .$$

Since $(1+\epsilon)\nu = \nu + \Delta/2 \leq \mu + \Delta/2$, by Chernoff bound,

$$\Pr[V_{j,s} \geq q(\mu + \Delta/2)] \leq \Pr[Y_1 + \cdots + Y_q \geq q(1+\epsilon)\nu]$$

$$\leq \exp\Big(\frac{-\epsilon^2 \nu q}{2+\epsilon}\Big) \leq \exp\Big(\frac{-\lambda M q}{9 \cdot N^{r-2}}\Big) \ . \tag{1}$$

Next, for each $i \leq q$, let Y_i^* be the Bernoulli random variable such that $Y_i^* = 1$ if and only if $U_{i,j^*} = s^*$. Again, the random variables $Y_1^*, \ldots, Y_q^*$ are independent and identically distributed, and $V_{j^*,s^*} = Y_1^* + \cdots + Y_q^*$. Let $\nu^* = \Pr[Y_1^* = 1] \geq \Delta + \mu$ and let $\epsilon^* = \frac{\Delta}{2(\mu+\Delta)}$. Then $0 < \epsilon^* < 1$. Moreover,

$$(\epsilon^*)^2 \nu^* \geq \frac{\Delta^2 q}{4(\mu+\Delta)} = \frac{\Delta^2/\mu}{4(1+\Delta/\mu)} \geq \frac{\Delta^2/\mu}{6} = \frac{\lambda M}{6 \cdot N^{r-2}} \ .$$

Since $(1-\epsilon^*)\nu^* \geq \Big(1 - \frac{\Delta}{2(\mu+\Delta)}\Big)(\Delta+\mu) = \mu + \Delta/2$, by Chernoff bound,

$$\begin{aligned} \Pr[V_{j^*,s^*} \leq q(\mu+\Delta/2)] &\leq \Pr[Y_1^* + \cdots + Y_q^* \leq q(1-\epsilon^*)\nu^*] \\ &\leq \exp\Big(\frac{-(\epsilon^*)^2 \nu^* q}{2}\Big) \leq \exp\Big(\frac{-\lambda M q}{12 \cdot N^{r-2}}\Big) \ . \end{aligned} \tag{2}$$

From Equation (1) and Equation (2), the chance that the adversary LD fails to recover Z_k is at most

$$\begin{aligned} \Pr[V_{j^*,s^*} \leq q(\mu+\Delta/2)] + \sum_{(j,s)\neq(j^*,s^*)} \Pr[V_{j,s} \geq q(\mu+\Delta/2)] \\ \leq \exp\Big(\tfrac{-\lambda M q}{12 \cdot N^{r-2}}\Big) + MN \cdot \exp\Big(\tfrac{-\lambda M q}{9 \cdot N^{r-2}}\Big) \ . \end{aligned}$$

5 Attacking FNR

In this section, we attack the Flexible Naor-Reingold (FNR) scheme proposed by Cisco [7], which is defined only for the boolean case.[12] It is based on Naor-Reingold generalization of Feistel networks [15], using a pairwise independent permutation and a boolean Feistel-based FPE scheme.

FNR CONSTRUCTION. Recall that a family $\mathcal{P}$ of permutations on $\{0,1\}^\ell$ is *pairwise independent* if for any $X, X', Y, Y' \in \{0,1\}^\ell$ such that $X \neq X'$ and $Y \neq Y'$,

$$\Pr_{\pi \xleftarrow{\$} \mathcal{P}}[(\pi(X) = Y) \wedge (\pi(X') = Y')] = \frac{1}{2^\ell(2^\ell - 1)} \ .$$

In FNR, the family $\mathcal{P}$ is instantiated as $\mathcal{B}_\ell$, the set of all pairs (B_0, B_1) such that B_0 is an invertible binary matrix of size $\ell \times \ell$, and B_1 is a binary vector of length ℓ. For each $\pi \in \mathcal{P}$, $\pi(X) = (B_0 \cdot X) \oplus B_1$, where the input X is viewed as a binary vector of length ℓ, (B_0, B_1) is the matrix representation of π, and the multiplication $B_0 \cdot X$ is in $\mathsf{GF}(2)$.

[12] While the FNR paper [7] mentions that the scheme can be used to encrypt credit-card numbers, it is unclear how this is possible, as the specific instantiation there only works for binary data.

$\underline{\mathsf{F.E}(K,T,X)}$
$(B_0, B_1, \widetilde{K}) \leftarrow K$
$(L,R) \leftarrow U \leftarrow (B_0 \cdot X) \oplus B_1$
For $i = 1$ to r do
  If $(i \bmod 2 = 1)$ then
    $L \leftarrow L \oplus F_i(\widetilde{K}, T, R)$
  Else $R \leftarrow R \oplus F_i(\widetilde{K}, T, L)$
$V \leftarrow (L,R)$; $Y \leftarrow B_0^{-1} \cdot (V \oplus B_1)$
Return Y

$\underline{\mathsf{F.D}(K,T,Y)}$
$(B_0, B_1, \widetilde{K}) \leftarrow K$
$(L,R) \leftarrow V \leftarrow (B_0 \cdot Y) \oplus B_1$
For $i = r$ to 1 do
  If $(i \bmod 2 = 1)$ then
    $L \leftarrow L \oplus F_i(\widetilde{K}, T, R)$
  Else $R \leftarrow R \oplus F_i(\widetilde{K}, T, L)$
$U \leftarrow (L,R)$; $X \leftarrow B_0^{-1} \cdot (U \oplus B_1)$
Return X

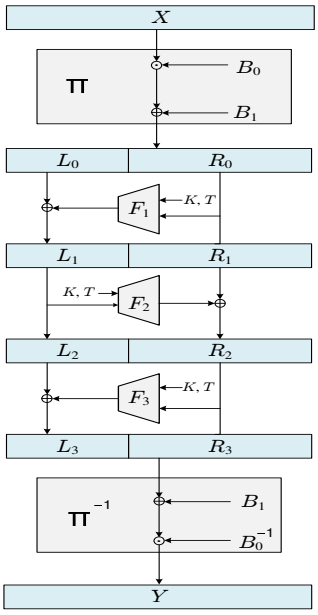

Fig. 5. Left: The code for the encryption and decryption algorithms of $\mathsf{F} = \mathbf{FNR}[r, m, n, \mathrm{PL}]$, where $\mathrm{PL} = (\mathcal{T}, \mathcal{K}, F_1, \ldots, F_r)$. In implementation, for $(L,R) \leftarrow U$, typically L is the leftmost m-bit substring of U, and R is the rightmost n-bit substring of U. However, in Cisco implementation, L and R are the strings obtained via the odd and even bits of U, respectively. **Right:** An illustration of encryption with $r = 3$ rounds, where $\odot$ denotes the matrix multiplication.

In an FNR scheme $\mathsf{F} = \mathbf{FNR}[r, m, n, \mathrm{PL}]$, the domain is $\{0,1\}^m \times \{0,1\}^n$. The parameter $\mathrm{PL} = (\mathcal{T}, \mathcal{K}, F_1, \ldots, F_r)$ specifies the tweak space $\mathcal{T}$ and a Feistel-based FPE scheme $\mathsf{F} = \mathbf{Feistel}[r, 2^m, 2^n, \oplus, \mathrm{PL}]$ as defined in Section 4. The key space is $\mathcal{B}_{m+n} \times \mathcal{K}$. On key $K = (B_0, B_1, \widetilde{K})$ and tweak T, to encrypt a message X, one first interprets (B_0, B_1) as a permutation $\pi : \{0,1\}^{m+n} \to \{0,1\}^{m+n}$, computes $U \leftarrow \pi(X)$ and $V \leftarrow \widetilde{\mathsf{F}}.\mathsf{E}(\widetilde{K}, T, U)$, and returns $\pi^{-1}(V)$. Decryption is defined likewise. The code of the encryption and decryption schemes of $\mathbf{FNR}[r, m, n, \mathrm{PL}]$ is given in Fig. 5. If the underlying Feistel-based FPE scheme is $\mathbf{Feistel}[r, 2^m, 2^n, \oplus]$ (meaning ideal round functions), then we write $\mathbf{FNR}[r, m, n]$ for the corresponding FNR scheme. For input length ℓ, the FNR specification only uses the $m = \lceil \ell/2 \rceil$ and $n = \ell - m$, meaning that the Feistel network is a (near)-balanced one. The suggested instantiation in [7] uses $r = 7$.

The FNR spec [7] specifies the round functions using AES. Again, using the standard assumption that AES is a good PRF, one can focus on attacking FNR schemes of ideal round functions, with small differences in the advantage.

The attack. We now attack the scheme $\mathbf{FNR}[r, m, n]$ scheme for an odd integer $r \geq 7$, with $|m - n| \leq 1$. This is exactly the setting specified by the FNR spec. While FNR also uses a Feistel network, at the first glance, it is unclear how

to use the ideas in Section 4, because the pairwise independent permutation in FNR will hide the pairwise bias described in Lemma 4. However, we will exploit the fact that the FNR scheme uses *the same* pairwise independent permutation across different tweaks.

Under our attack, there are $\tau = \lceil 2^n(\ln(2)\cdot n + 3)\rceil$ known messages $X_1, \ldots, X_\tau$ sampled uniformly without replacement from $\{0,1\}^{m+n}$, and there are p targets $Z_1, \ldots, Z_p$. The adversary is given the encryption of those $\tau + p$ messages under q tweaks $T_1, \ldots, T_q$, for an appropriately large q, and the auxiliary information is $(X_1, \ldots, X_\tau, p, q)$. From the distinctness requirement, these $\tau + p$ messages must be distinct. We have no other restriction on the number p of targets, so potentially p can be as large as $2^{m+n} - \tau$. Our attack will recover all of $Z_1, \ldots, Z_p$, meaning $d = p$. The number of examples Q is $(\tau + p)q$, and the number of examples per target q_t is $(\tau/p + 1)q$.

We formalize the attack via the message-recovery framework, by specifying a class $\mathsf{SC2}_{p,q,\theta}$ of samplers, and then giving a lower bound on the mr-advantage of the attack for any sampler in this class. First, let $\mathsf{DC2}_{p,q,\theta}$ be the class of all algorithms D that outputs q distinct tweaks $T_1, \ldots, T_q \in \{0,1\}^*$, and distinct $X_1, \ldots, X_\tau, Z_1, \ldots, Z_p \in \{0,1\}^{m+n}$ such that (1) $X_1, \ldots, X_\tau$ are sampled uniformly without replacement from $\{0,1\}^{m+n}$, and (2) given $X_1, \ldots, X_\tau, T_1, \ldots, T_q$, for any fixed $Z_1^*, \ldots, Z_p^*$, the conditional probability that $Z_1 = Z_1^*, \ldots, Z_p = Z_p^*$ is at most $2^{-\theta}$. To any such D, we associate the sampler

$$\begin{array}{l}\underline{\text{Sampler } \mathsf{XS}[\mathsf{D}]}\\ (T_1, \ldots, T_q, X_1, \ldots, X_\tau, Z_1, \ldots, Z_p) \leftarrow\!\!\$\ \mathsf{D}\\ a \leftarrow (X_1, \ldots, X_\tau, p, q)\\ \text{Return } \big(\{(T_i, X_j), (T_i, Z_k) \mid i \le q, j \le \tau, k \le p\}, Z_1, \ldots, Z_p, a\big)\end{array}$$

The sampler above return the pairs (T_i, X_j) and (T_i, Z_k) for every $i \le q, j \le \tau$, and $k \le p$, where the targets are $Z_1, \ldots, Z_p$. Let $\mathsf{SC2}_{p,q,\theta} = \{\mathsf{XS}[\mathsf{D}] \mid \mathsf{D} \in \mathsf{DC2}_{p,q,\theta}\}$. The Full-message Differential (FD) attack, given in Fig. 6, can recover all targets $Z_1, \ldots, Z_p$ in $O(pq\tau)$ time. Theorem 6 below gives a lower bound on the mr-advantage of LD; the proof is postponed further below. The bound is illustrated in Fig. 7.

Theorem 6. *Let $m, n \ge 3$ and $q \ge 1$ be integers such that $|m - n| \le 1$, and let $r \ge 7$ be an odd integer. Let $\mathsf{F} = \mathbf{FNR}[r, m, n]$. Let $\lambda = \left(1 - \frac{1}{2^n - 1}\right)^2 \left(1 - \frac{1}{2^{m+n}}\right)$. Then for any $\theta \ge 0$ and for any sampler XS in the class $\mathsf{SC2}_{p,q,\theta}$,*

$$\begin{aligned}\mathbf{Adv}^{\mathsf{mr}}_{\mathsf{F},\mathsf{XS}}(\mathsf{FD}) \ge 0.95 &- 2^{m+n}p \cdot \exp\Big(\frac{-q}{32 \cdot 2^{3(m+n)}}\Big) - 2^{m+n}p \cdot \exp\Big(\frac{-q}{48 \cdot 2^{3(m+n)}}\Big)\\ &- 2^{m+n}p \cdot \exp\Big(\frac{-\lambda q}{9 \cdot 2^{n+(r-2)m}}\Big) - p \cdot \exp\Big(\frac{-\lambda q}{12 \cdot 2^{n+(r-2)m}}\Big) - 2^{-\theta}.\end{aligned}$$

Ideas of the attack. For a random variable $W \in \{0,1\}^{m+n}$, we say that it has a *singular* distribution if there is exactly one string $Z \in \{0,1\}^{m+n}$ such

Adversary $\mathsf{FD}(\{(T_i, C_{i,j}), (T_i, C'_{i,k}) \mid i \le q, j \le \tau, k \le p\}, a)$
$(X_1, \ldots, X_\tau, p, q) \leftarrow a$; $\mu \leftarrow 1/2^{m+n}$; $\Delta \leftarrow \frac{1}{2 \cdot 2^{2(m+n)}}$
For $k \leftarrow 1$ to p do // Recover target Z_k
  For $j \in \{1, \ldots, \tau\}, s \in \{0,1\}^{m+n}$ do $V_{j,s} \leftarrow 0$
  For $i \leftarrow 1$ to q, $j \leftarrow 1$ to τ do $s \leftarrow C_{i,j} \oplus C'_{i,k}$; $V_{j,s} \leftarrow V_{j,s} + 1$
  Find smallest j^* s.t. there is only one $s \in \{0,1\}^{m+n}$ with $V_{j^*,s} \le q(\mu + \Delta/2)$.
  Let $V_{j^*,s^*} = \max\{V_{j^*,s} \mid s \in \{0,1\}^{m+n}\}$; $Z_k \leftarrow s^* \oplus X_{j^*}$
Return $(Z_1, \ldots, Z_p)$

Fig. 6. The Full-message Differential attack.

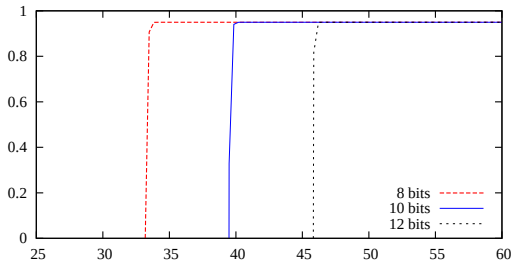

Fig. 7. The mr advantage of the Full-message Differential attack on $\mathbf{FNR}[r,n,n]$ for $r = 7$ and $n = 4, 5, 6$. This is the balanced setting $m = n$. The x-axis shows the log, base 2, of the number q of tweaks (which is also roughly q_t, the number of ciphertexts per recovered target), and the y-axis shows $\mathbf{Adv}^{\mathsf{mr}}_{\mathbf{FNR}[r,n,n],\mathsf{XS}}(\mathsf{FD})$, for XS that outputs $\tau = \lceil 2^n(\ln(2) \cdot n + 3) \rceil$ known messages and $p = 2^{2n} - \tau$ targets; those 2^{2n} messages are sampled uniformly without replacement from $\{0,1\}^{2n}$.

that $\Pr[W = Z] \le 1/2^{m+n}$; otherwise the distribution is *non-singular*. Let $\pi = (B_0, B_1)$ be the pairwise independent permutation in the key of the FNR scheme. Suppose that one encrypts distinct messages X and X' on a tweak T. Then the strings $Y \leftarrow \pi(X)$ and $Y' \leftarrow \pi(X')$ become inputs to a near-balanced, boolean Feistel network, and let U and U' be the corresponding outputs of the Feistel network. Our attack is based on the following observation that is formalized in Lemma 7 below; see Appendix B for the proof. Specifically, if Y and Y' have the different right segments then the distribution of $U \oplus U'$ is non-singular; in fact, there are 2^m values $Z \in \{0,1\}^{m+n}$ such that $\Pr[U \oplus U' = Z] \le 1/2^{m+n}$. Let C and C' be the ciphertexts of Y and Y' under the FNR scheme, respectively. Then $C \leftarrow \pi^{-1}(U)$ and $C' \leftarrow \pi^{-1}(U')$, and $C \oplus C' = B_0^{-1} \cdot (U \oplus U')$. Thus the distribution of $C \oplus C'$ is also non-singular.

In contrast, suppose that Y and Y' have the same right segments. Then $\Pr[U \oplus U' = Z]$ is significantly larger than $1/2^{m+n}$ for every $Z \in \{0,1\}^{m+n} \backslash \{0^{m+n}\}$, and thus the distribution of $U \oplus U'$, and also that of $C \oplus C'$, are singular in this case. Moreover, the distribution of $U \oplus U'$ peaks at

$Y \oplus Y' = B_0 \cdot (X \oplus X')$, and consequently, the distribution of $C \oplus C'$ peaks at $B_0^{-1} \cdot B_0 \cdot (X \oplus X') = X \oplus X'$.

Lemma 7. *Let $r \geq 7$ be an odd integer and let $m, n \geq 2$ be integers such that $|m - n| \leq 1$. Let $\mathsf{F} = \mathbf{Feistel}[r, 2^m, 2^n, \oplus]$. Fix distinct $X, X' \in \{0,1\}^{m+n}$, a tweak $T \in \mathsf{F.Twk}$. Pick $K \leftarrow\!\!\$\ \mathsf{F.Keys}$. For each integer t, let X_t and X'_t be the the round-t output of X and X' under $\mathsf{F}(K, T, \cdot)$, respectively. Then for any odd integer $t \geq 7$,*

(a) If X and X' have different right segments then for any non-zero $Z \in \{0,1\}^{m+n}$,

$$\Pr[X_t \oplus X'_t = Z] = \frac{1}{2^{m+n}} \text{ if } \mathbf{Right}(Z) = 0^n \ ,$$
$$\Pr[X_t \oplus X'_t = Z] \geq \frac{1}{2^{m+n}} + \frac{1}{2 \cdot 2^{2(m+n)}} \text{ otherwise } .$$

(b) If X and X' have the same right segments then for any non-zero $Z \in \{0,1\}^{m+n}$,

$$\Pr[X_t \oplus X'_t = Z] \geq \frac{1}{2^{m+n}} + \frac{1}{2 \cdot 2^{2(m+n)}} \ .$$

Moreover,

$$\Pr[X_t \oplus X'_t = Z] \leq \frac{1}{2^{m+n} - 1} + \frac{1}{(2^m - 1)2^{(t-1)(m+n)/2}} \text{ if } Z \neq X \oplus X',$$
$$\Pr[X_t \oplus X'_t = Z] \geq \frac{1}{2^{m+n} - 1} + \frac{1 - 1/(2^m - 1)}{2^n \cdot 2^{(t-1)m/2}} \text{ otherwise } .$$

The probabilities above are taken over a sampling $K \leftarrow\!\!\$\ \mathsf{F.Keys}$.

Based on the observation above, we can attack the FNR scheme as follows. The adversary receives the encryptions of known messages $X_1, \ldots, X_\tau$ and targets $Z_1, \ldots, Z_p$, under tweaks $T_1, \ldots, T_q$. Fix $k \leq p$; we now explain how to recover Z_k. Let $C_{i,j}$ and $C'_{i,k}$ be the ciphertexts of X_j and Z_k under tweak T_i, respectively. To recover a target Z_k, for each $j \leq \tau$, we plot the frequency histogram for the values $C_{i,j} \oplus C'_{i,k}$, for every $i = 1, \ldots, q$, and call it the histogram of X_j. From the observation above, if $\pi(X_j)$ and $\pi(Z_k)$ have different right segments and q is big enough then the histogram for X_j is *non-singular*, meaning that it has multiple short columns, relative to the height $q/2^{m+n}$. In contrast, if $\pi(X_j)$ and $\pi(Z_k)$ have the same right segments then the histogram for X_j is singular, containing exactly one short column (of height 0). Moreover, in this case, the tallest column corresponds to the value $X_j \oplus Z_k$.

Since $X_1, \ldots, X_\tau$ are sampled uniformly without replacement from $\{0,1\}^{m+n}$ and π is a permutation on $\{0,1\}^{m+n}$, the strings $Y_1 \leftarrow \pi(X_1), \ldots, Y_\tau \leftarrow \pi(X_\tau)$ are also sampled uniformly without replacement from $\{0,1\}^{m+n}$. From the Coupon Collector's problem (Lemma 2), $\{\mathbf{Right}(Y_1), \ldots, \mathbf{Right}(Y_\tau)\} = \{0,1\}^n$ with probability at least 0.95. Hence there must be some j^* such that Y_{j^*} and $\pi(Z_k)$ have the same right segment. We can find such a j^* by checking if its

histogram is singular. Let s^* be the value for the tallest column in the histogram of X_{j^*}. We then can recover Z_k by way of $Z_k \leftarrow s^* \oplus X_{j^*}$.

Proof of Theorem 6. First we show that $\mathbf{Adv}^{\mathsf{mg}}_{\mathsf{XS}} \le 2^{-\theta}$. Consider an arbitrary simulator $\mathcal{S}$. To win the game, $\mathcal{S}$ must guess all targets, given the tweaks and the auxiliary information. From the definition of θ, the chance that the simulator's guess is correct is at most $2^{-\theta}$. Next, we show that

$$\begin{aligned}&\Pr[\mathbf{G}^{\mathsf{mr}}_{\mathsf{F},\mathsf{XS}}(\mathsf{FD})]\\ &\ge 0.95 - 2^{m+n}p \cdot \exp\Big(\frac{-q}{32 \cdot 2^{3(m+n)}}\Big) - 2^{m+n}p \cdot \exp\Big(\frac{-q}{48 \cdot 2^{3(m+n)}}\Big)\\ &\quad -2^{m+n}p \cdot \exp\Big(\frac{-\lambda q}{9 \cdot 2^{n+(r-2)m}}\Big) - p \cdot \exp\Big(\frac{-\lambda q}{12 \cdot 2^{n+(r-2)m}}\Big)\ .\end{aligned}$$

Let $Y \leftarrow \pi(X_1), \ldots, Y_\tau \leftarrow \pi(X_\tau)$. Since $X_1, \ldots, X_\tau$ are sampled uniformly without replacement from $\{0,1\}^{m+n}$ and π is a permutation on $\{0,1\}^{m+n}$, the strings $Y_1, \ldots, Y_\tau$ are also sampled uniformly without replacement from $\{0,1\}^{m+n}$. From the Coupon Collector's problem, $\{\mathbf{Right}(Y_1), \ldots, \mathbf{Right}(Y_\tau)\} = \{0,1\}^n$, with probability at least 0.95. By union bound, it suffices to prove that for any $k \le p$, the FD attack fails to recover the target Z_k with probability at most

$$\begin{aligned}&2^{m+n} \cdot \exp\Big(\frac{-q}{32 \cdot 2^{3(m+n)}}\Big) + 2^{m+n} \cdot \exp\Big(\frac{-q}{48 \cdot 2^{3(m+n)}}\Big)\\ &+2^{m+n} \cdot \exp\Big(\frac{-\lambda q}{9 \cdot 2^{n+(r-2)m}}\Big) + \exp\Big(\frac{-\lambda q}{12 \cdot 2^{n+(r-2)m}}\Big)\ .\end{aligned}$$

Let $C_{i,j}$ and $C'_{i,k}$ be the ciphertexts for known messages X_i and target Z_k under tweak T_i, respectively. Let $B_{j,i,s}$ be the Bernoulli random variable such that $B_{i,j,s} = 1$ if and only if $C_{i,j} \oplus C'_{i,k} = s$. Now in the histogram for X_j, the height of the column for each value s is $V_{j,s} = B_{1,j,s} + \cdots + B_{q,j,s}$. Note that for each fixed (j,s), the random variables $B_{1,j,s}, \ldots, B_{q,j,s}$ are independent and identically distributed. Let $\mu \leftarrow 1/2^{m+n}$ and $\Delta \leftarrow \frac{1}{2 \cdot 2^{2(m+n)}}$. From Chernoff bound,

(i) For every (j,s), if $\Pr[B_{1,j,s} = 1] \le \mu$ then $V_{j,s} \ge q(\mu + \Delta/2)$ with probability at most $\exp\Big(\frac{-q}{32 \cdot 2^{3(m+n)}}\Big)$. That is, a supposedly short column is likely to remain short.

(ii) For every (j,s), if $\Pr[B_{1,j,s} = 1] \ge \mu + \Delta$, we have $V_{j,s} \le q(\mu + \Delta/2)$ with probability at most $\exp\Big(\frac{-q}{48 \cdot 2^{3(m+n)}}\Big)$. That is, a supposedly tall column will be likely to remain tall.

Now, consider j such that $\pi(X_j)$ and $\pi(Z_k)$ have different right segments. Since $X_j \ne Z_k$ and FNR is a permutation, the histogram for X_j will surely have one column of height 0, namely the column corresponding to $\pi(0^{m+n})$. To correctly identify the histogram as non-singular, we need one more supposedly short column of this histogram to remain short. From the claim (i) above and from Lemma 7, this happens for every such j with probability at least

$$1 - \tau \cdot \exp\Big(\frac{-q}{32 \cdot 2^{3(m+n)}}\Big) \ge 1 - 2^{m+n} \cdot \exp\Big(\frac{-q}{32 \cdot 2^{3(m+n)}}\Big)\ .$$

Next, consider the smallest j^* such that $\pi(X_{j^*})$ and $\pi(Z_k)$ have the same right segment. Since $X_{j^*} \neq Z_k$ and FNR is a permutation, the histogram for X_{j^*} will surely have one column of height 0, namely the column corresponding to $\pi(0^{m+n})$. To correctly identify the histogram as singular, we need every supposedly tall column of this histogram to remain tall. From the claim (ii) above and from Lemma 7, this happens with probability at least

$$1 - 2^{m+n} \cdot \exp\Big(\frac{-q}{48 \cdot 2^{3(m+n)}}\Big) \ .$$

By a union bound, we can realize j^* via checking the singularity of histograms with probability at least

$$1 - 2^{m+n} \cdot \exp\Big(\frac{-q}{32 \cdot 2^{3(m+n)}}\Big) - 2^{m+n} \cdot \exp\Big(\frac{-q}{48 \cdot 2^{3(m+n)}}\Big) \ . \tag{3}$$

Now, once we find j^*, we need to ensure that the peak column indeed corresponds to the value $X_{j^*} \oplus Z_k$. Let $\mu^* = \frac{1}{2^{m+n}-1} + \frac{1/(2^m-1)}{2^{(r-1)(m+n)/2}}$ and $\Delta^* = \frac{1-1/(2^m-2)}{2^n \cdot 2^{(r-1)m/2}}$. From Chernoff bound and Lemma 7,

(iii) For every $s \neq Z_k \oplus X_{j^*}$, $\Pr[B_{1,j^*,s} = 1] \leq \mu^*$, and thus the probability that $V_{j^*,s} \geq q(\mu^* + \Delta^*/2)$ is at most $\exp\Big(\frac{-\lambda q}{9 \cdot 2^{n+(r-2)m}}\Big)$. That is, it is unlikely that the column corresponding to s is the peak, as it remains lower than $q(\mu^* + \Delta^*/2)$.

(iv) For $s^* = Z_k \oplus X_{j^*}$, $\Pr[B_{1,j^*,s^*} = 1] \geq \mu^* + \Delta^*$, and thus $V_{j^*,s^*} \leq q(\mu^* + \Delta^*/2)$ with probability at most $\exp\Big(\frac{-\lambda q}{12 \cdot 2^{n+(r-2)m}}\Big)$. That is, the column corresponding to $Z_k \oplus X_{j^*}$ is likely to be the peak, as it remains higher than $q(\mu^* + \Delta^*/2)$.

From (iii) and (iv), the chance that in the histogram of X_{j^*}, the peak column indeed corresponds to $X_{j^*} \oplus Z_k$ is at least

$$1 - 2^{m+n} \cdot \exp\Big(\frac{-\lambda q}{9 \cdot 2^{n+(r-2)m}}\Big) - \exp\Big(\frac{-\lambda q}{12 \cdot 2^{n+(r-2)m}}\Big) \ . \tag{4}$$

From Equation (3) and Equation (4), the chance that the attack can recover the target Z_k is at least

$$\begin{aligned} &1 - 2^{m+n} \cdot \exp\Big(\frac{-q}{32 \cdot 2^{3(m+n)}}\Big) - 2^{m+n} \cdot \exp\Big(\frac{-q}{48 \cdot 2^{3(m+n)}}\Big) \\ &-2^{m+n} \cdot \exp\Big(\frac{-\lambda q}{9 \cdot 2^{n+(r-2)m}}\Big) - \exp\Big(\frac{-\lambda q}{12 \cdot 2^{n+(r-2)m}}\Big) \ . \end{aligned}$$

This completes the proof.

6 Attacking DTP

In this section, we will attack the DTP scheme, by Protegrity Corp. [12], which resembles the seminal FPE construction by Brightwell and Smith [6].

$\underline{\mathsf{F.E}(K,T,X)}$	$\underline{\mathsf{F.D}(K,T,Y)}$
$x_1 \cdots x_m \leftarrow X$; $T_1 \leftarrow T$; $t \leftarrow \lfloor m/r \rfloor$	$y_1 \cdots y_m \leftarrow Y$; $T_1 \leftarrow T$; $t \leftarrow \lfloor m/r \rfloor$
For $i = 1$ to t do	For $i = 1$ to t do
$\quad z_1 \cdots z_n \leftarrow F_K(T_i)$; $k \leftarrow (i-1)r$	$\quad z_1 \cdots z_n \leftarrow F_K(T_i)$; $k \leftarrow (i-1)r$
$\quad$ For $j = 1$ to r do	$\quad$ For $j = 1$ to r do
$\quad\quad y_{k+j} \leftarrow (x_{k+j} + z_j) \bmod d$	$\quad\quad x_{k+j} \leftarrow (y_{k+j} - z_j) \bmod d$
$\quad T_{i+1} \leftarrow z_{r+1} \cdots z_n x_{k+1} \cdots x_{k+r}$	$\quad T_{i+1} \leftarrow z_{r+1} \cdots z_n x_{k+1} \cdots x_{k+r}$
// Encrypt the trailing digits	// Decrypt the trailing digits
$z_1 \cdots z_n \leftarrow F_K(T_{t+1})$	$z_1 \cdots z_n \leftarrow F_K(T_{t+1})$
For $j = 1$ to $(m \bmod r)$ do	For $j = 1$ to $(m \bmod r)$ do
$\quad\quad y_{tr+j} \leftarrow (x_{tr+j} + z_j) \bmod d$	$\quad\quad x_{tr+j} \leftarrow (y_{tr+j} - z_j) \bmod d$
Return $y_1 \cdots y_m$	Return $x_1 \cdots x_m$

Fig. 8. Code for the encryption and decryption algorithms of $\mathsf{F} = \mathbf{DTP}[r, d, D, m, n, \mathrm{PL}]$, where $\mathrm{PL} = (\mathcal{K}, F)$.

DTP CONSTRUCTION. The DTP scheme has several variants, but here we only consider the simplest and also the most efficient one. Under this version, it requires that each time we encrypt a message, we need to pick a fresh random tweak. Thus in this setting, tweaks serve the same role as initialization vectors in traditional modes of encryption like CBC.

The scheme $\mathsf{F} = \mathbf{DTP}[r, d, D, m, n, \mathrm{PL}]$ has message space $\mathbb{Z}_d^m$ and tweak space $\mathbb{Z}_D^n$, with $d \le D$ and $n \ge r$. The parameter $\mathrm{PL} = (\mathcal{K}, F)$ specifies the key space $\mathcal{K}$ and the round function $F : \mathcal{K} \times \mathbb{Z}_D^n \to \mathbb{Z}_D^n$. For example, if we want to encrypt credit-card numbers (CCNs) then $m = 16$, and there are two possible values for d:

(i) Conventionally, one views CCNs as a sequence of decimal digits, and thus $d = 10$.
(ii) Protegrity prefers to interpret CCNs as a sequence of (case-senstive) alphanumeric characters for seemingly better security, and thus $d = 62$.

Under the specification in [12], one then instantiates the round function F from AES, interpreting $\{0,1\}^{128}$ as $\mathbb{Z}_{256}^{16}$ (meaning $n = 16$ and $D = 256$). The code for the encryption and decryption of F is given in Fig. 8. The DTP specification always uses $D = 256$ if $d \le 256$, and $D = 2^{16}$ if d is bigger. The parameter r specifies how many input characters that one encrypts per one call to the round function F. Initially, Protegrity used $r = 1$; this version is known internally as DTP-1. Eventually, they moved to $r = 3$ for faster speed, and also claimed better security; this is the current version, known as DTP-2.

If we consider an ideal round function then $\mathcal{K}$ is the set of all functions $G : \mathbb{Z}_D^n \to \mathbb{Z}_D^n$, and $F_K(\cdot)$ is defined as the function $G(\cdot)$ that the key K encodes. We write $\mathbf{DTP}[r, d, D, m, n]$ to denote the DTP construction of this particular choice of $\mathrm{PL} = (\mathcal{K}, F)$. As mentioned above, since the DTP spec instantiates

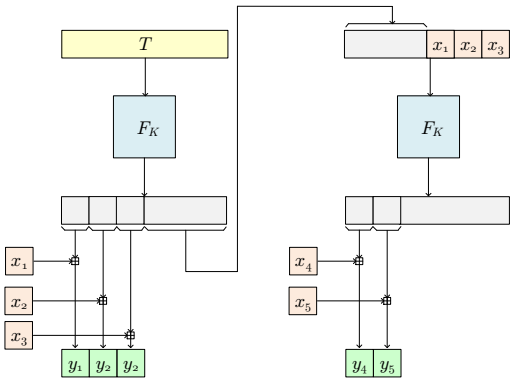

Fig. 9. Illustration of the encryption scheme of $\mathsf{F} = \mathbf{DTP}[r, d, D, m, n, \mathrm{PL}]$, where $\mathrm{PL} = (\mathcal{K}, F)$, for $r = 3$ and $m = 5$, and $\boxplus$ means the addition in mod d.

the round function via AES, using the standard assumption that AES is a good PRF, one can focus on attacking DTP schemes of ideal round functions, with small differences in the advantage.

THE ATTACK. We now give an attack on a general $\mathbf{DTP}[r, d, D, m, n]$ scheme in which d is *not* a divisor of D. Many applications of DTP use $d = 10$ or $d = 62$ (for examples, encrypting credit-card numbers, social-security numbers, or PINs), and in that case, $D = 256$, falling into our setting. In this attack, we consider only a single target Z. There is no known message, and the auxiliary information is null. The adversary is given the encryption of Z under tweaks $T_1, \ldots, T_q$, for an appropriately large q. The number Q of ciphertexts is q, and so is the number of ciphertexts per recovered target. We assume that Z is uniformly random, independent of the tweaks, so that the message-guessing advantage is low.

Formally, let $\mathsf{DC3}_q$ be the class of all algorithms D that outputs distinct tweaks $T_1, \ldots, T_q \in (\mathbb{Z}_D)^n$. To any such D, we associate the following sampler $\mathsf{XS}[\mathsf{D}]$

Sampler $\mathsf{XS}[\mathsf{D}]$
$(T_1, \ldots, T_q) \leftarrow\!\!\$\ \mathsf{D};\ a \leftarrow \bot;\ Z \leftarrow\!\!\$\ (\mathbb{Z}_d)^m$
Return $((T_1, Z), \ldots, (T_q, Z), Z, a)$

The sampler above runs D to generate the tweaks, and then samples a uniformly random target. Define $\mathsf{SC3}_q = \{\mathsf{XS}[\mathsf{D}] \mid \mathsf{D} \in \mathsf{DC3}_q\}$. Since the target is uniformly random and the auxiliary information is null, one would expect that the adversary has low mr-advantage, even if q is big. However, our Digit-wise Differential (DD) attack, given in Fig. 10, will recover the target message for any sampler in $\mathsf{SC3}_q$ within $O(md \log(d) + qm)$ time. Theorem 8 below gives a lower bound on the mr-advantage of DD; the proof is in Appendix C. The bound is illustrated in Fig. 11.

Adversary $\mathsf{DD}((T_1, C_1), \ldots, (T_q, C_q), a)$

For $i \leftarrow 1$ to m do
  For $k \in \mathbb{Z}_d$ do $V_k \leftarrow 0$
  For $j \leftarrow 1$ to q do $c_1 \cdots c_m \leftarrow C_j$; $V_{c_i} \leftarrow V_{c_i} + 1$
  $r \leftarrow D \bmod d$
  Find the r largest numbers $V_{s_1}, \ldots, V_{s_r}$ in $\{V_k \mid k \in \mathbb{Z}_d\}$
  Find $z_i \in \mathbb{Z}_d$ such that $\{s_j \mid 1 \le j \le r\} = \{(z_i + j) \bmod d \mid 1 \le j \le r\}$
$Z \leftarrow z_1 \cdots z_m$; Return Z

Fig. 10. The Digit-wise Differential attack.

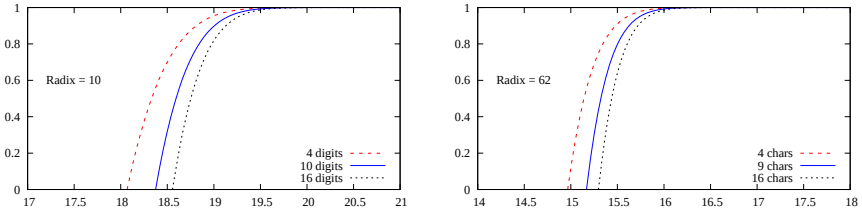

Fig. 11. The mr advantage of the Digit-wise Differential attack on $\mathbf{DTP}[3, 10, 256, m, 16]$ (left) and $\mathbf{DTP}[3, 62, 256, m, 16]$ (right) for $m = 4, 9, 16$. These are parameter choices for PINs, social security numbers, and credit-card numbers. The x-axis shows the log, base 2, of the number q of ciphertexts, and the y-axis shows $\mathbf{Adv}^{\mathrm{mr}}_{\mathbf{DTP}[3,d,256,m,16],\mathsf{XS}}(\mathsf{DD})$ for $\mathsf{XS} \in \mathsf{SC3}_q$.

Theorem 8. *Let $D > d > 1$ be integers such that d is not a divisor of D. Let $m, n, r \ge 1$ be integers such that $n \ge r$, and let $\mathsf{F} = \mathbf{DTP}[r, d, D, m, n]$. Let $s = D \bmod d$. Then for any sampler XS in $\mathsf{SC3}_q$,*

$$\mathbf{Adv}^{\mathrm{mr}}_{\mathsf{F},\mathsf{XS}}(\mathsf{DD}) \ge 1 - \frac{(q \cdot \lceil m/r \rceil)^2}{2 \cdot D^{n-r}} - ms \cdot \exp\Big(\frac{-q(d-s)^2}{2Dd(D+d-s)}\Big) - m(d-s) \cdot \exp\Big(\frac{-qs^2}{3Dd(D-s)}\Big) - \frac{1}{d^m} \ .$$

Ideas of the attack. For simplicity, let us start with the special important case $d = 10$ and $D = 256$. Let $Z = z_1 \cdots z_m$, where each z_i is a number in $\{0, \ldots, 9\}$. For simplicity, assume that the $q \cdot \lceil m/r \rceil$ inputs to F are distinct, so that the outputs of F are independent, which holds with high probability. We now explain how the attack can recover, say the first digit z_1 of the target Z, but the same idea works for any digit z_i of Z. The way the encryption works is to pick a random number $B \xleftarrow{\$} \{0, \ldots, 255\}$, and then outputs $c_1 \leftarrow z_1 + (B \bmod 10)$ as the first digit of the ciphertext. The problem here is that $B \bmod 10$ is not uniformly distributed in $\{0, 1, \ldots, 9\}$. In fact, for $a \in \{0, 1, \ldots, 9\}$, the probability

Radix d	**PINs** ($m = 4$)	**SSNs** ($m = 9$)	**CCNs** ($m = 16$)
10	$460,000$	$525,000$	$575,000$
62	$46,000$	$51,000$	$53,000$

Table 4. Comparison of security of DTP-2 over the choice of the radix d, on PINs, social security numbers, and credit-card numbers. The first column shows the value of d. The other columns show the estimated number of ciphertexts needed for our attack to achieve advantage 0.9 as suggested by Theorem 8.

that $B = a$ is exactly $\frac{\lceil 256/10 \rceil}{256} = \frac{26}{256}$ if $a < 6$, and this probability however is only $\frac{\lfloor 256/10 \rfloor}{256} = \frac{25}{256}$ otherwise. Hence for any fixed number $z_1 \in \{0, 1, \ldots, 9\}$ and any number $a \in \{0, 1, \ldots, 9\}$, the probability that $c_1 \leftarrow z_1 + (B \bmod 10)$ is a is exactly $\frac{26}{256}$ if $a \in \{z_1 \bmod 10, z_1 + 1 \bmod 10, \ldots, z_1 + 5 \bmod 10\}$, and is $\frac{25}{256}$ otherwise. Thus if we encrypt the target Z with a large enough number of times and plot the frequency histogram of the first digit of the ciphertexts, then what we obtain is a 10-column histogram, with 6 tall columns and 4 short ones. These 6 tall columns will be consecutive (possibly with a wrap-around), and the first one corresponds to the value z_1.

Now suppose that we want to deal with generic D and d, but d is not a divisor of D. Let $Z = z_1 \cdots z_m$, where each z_i is a number in $\mathbb{Z}_d$. Consider, say the first digit z_1 of Z. The encryption works by picking a random number $B \leftarrow\!\!\$\, \mathbb{Z}_D$ and then outputs $c_1 \leftarrow z_1 + (B \bmod d)$ as the first digit of the ciphertext. Again because d is not a divisor of D, the random variable $B \bmod d$ is not uniformly distributed in $\mathbb{Z}_d$. In fact, for $a \in \mathbb{Z}_d$, the probability that $B = a$ is exactly $\frac{\lceil D/d \rceil}{D}$ if $a < D \bmod d$, and this probability however is only $\frac{\lfloor D/d \rfloor}{D}$ otherwise. By the same argument as the special case above, if we encrypt the target Z with a large enough number of times and plot the frequency histogram of the first digit of the ciphertexts, then what we obtain is a histogram, with $D \bmod d$ tall columns. These tall columns will be consecutive (possibly with a wrap-around), and the first one corresponds to the value z_1.

Discussion. As Theorem 8 suggests, the security of DTP-2 (namely $r = 3$) is not better than that of DTP-1 (namely $r = 1$). Moreover, Protegrity's decision to prefer $d = 62$ over $d = 10$ actually makes security *worse*. As shown in Table 4, if one interprets a CCN as a sequence of 16 decimal digits, then one would need to obtain roughly $575,000$ ciphertexts to recover a CCN with advantage at least 0.9. In contrast, if one interprets a CCN as a sequence of 16 alphanumeric characters, then one would only need about $53,000$ ciphertexts to recover a CCN with advantage at least 0.9.

Experiments. We implement our Digit-wise Differential attack in C++ and evaluate its message-recovery rate against both DTP-1 and DTP-2, for domains $\mathbb{Z}_d^m$, with $m \in \{4, 9, 16\}$ and $d \in \{10, 62\}$. (For DTP-1, we only use $d = 10$.) Each experiment for domain $\mathbb{Z}_d^m$ was run using m threads in a server of Intel(R) Xeon(R) CPU E5-2699 v3 2.30GHz CPU and 256 GB RAM. For each setting,

Domain	Number of tweaks, q	Recovery rate	Time (ms)	Number of tweaks, q	Recovery rate	Time (ms)
$\mathbb{Z}_{10}^4$	2^{18}	100%	2.9	2^{17}	98%	1
$\mathbb{Z}_{10}^9$		100%	3		91%	1.49
$\mathbb{Z}_{10}^{16}$		100%	3.5		83%	1.87

Table 5. Empirical results of the Digit-wise Differential attack on DTP-1. For each domain (shown in the first column), we run experiments with two values of q (the number of tweaks) as indicated in the second and fifth columns. The recovery rates corresponding to these two values of q are given in the third and sixth columns, respectively. Finally, the average running time (in milliseconds) of each experiment is given in the fourth and seventh columns.

Domain	Number of tweaks, q	Recovery rate	Time (ms)	Number of tweaks, q	Recovery rate	Time (ms)
$\mathbb{Z}_{10}^4$	2^{18}	100%	3	2^{17}	95%	1
$\mathbb{Z}_{10}^9$		100%	3.08		90%	1.53
$\mathbb{Z}_{10}^{16}$		100%	3.58		83%	1.97
$\mathbb{Z}_{62}^4$	2^{16}	100%	0.01	2^{15}	91%	0.01
$\mathbb{Z}_{62}^9$		100%	1.03		78%	0.02
$\mathbb{Z}_{62}^{16}$		100%	1.17		68%	1

Table 6. Empirical results of the Digit-wise Differential attack on DTP-2.

we run our attack for several choices of q (the number of tweaks), each for 100 trials, and report the average running time and the empirical recovery rate.

Our experimental results for DTP-1, given in Table 5, are quite consistent with Theorem 8. For example, for domain $\mathbb{Z}_{10}^{16}$ (namely CCNs), theoretically one would need $q = 2^{19}$ tweaks to recover the target with probability nearly 1, and our experiments confirm that using $q = 2^{19}$ indeed gives 100% recovery rate. However, empirically, we find that $q = 2^{18}$ is enough to achieve 100% recovery rate, and each trial takes just 3.5 ms on average. If one instead uses $q = 2^{17}$, the recovery rate drops to 83%.

The experimental results for DTP-2 are given in Table 6, confirming the theoretical observations in Table 4: (1) DTP-2 is just as insecure as DTP-1, and (2) Using radix $d = 62$ instead of $d = 10$ exacerbates the insecurity: for example, for $\mathbb{Z}_{62}^{16}$ (namely CCNs), using $q = 2^{15}$ is already enough to achieve 68% recovery rate, and using $q = 2^{16}$ results in 100% recovery rate.

Acknowledgments

We thank Mihir Bellare and all anonymous reviewers for insightful feedback. We also thank Michael Maloney and Clyde Williamson of Protegrity Corp. for providing the information of the DTP scheme.

Viet Tung Hoang was supported by NSF grants CICI-1738912 and CRII-1755539. Stefano Tessaro was supported by NSF grants CNS-1553758 (CAREER), CNS-1423566, CNS-1719146, CNS-1528178, and IIS-1528041, and by a Sloan Research Fellowship. Ni Trieu was supported by NSF award #1617197.

References

1. M. Bellare, V. T. Hoang, and S. Tessaro. Message-recovery attacks on feistel-based format preserving encryption. In *ACM CCS 16*, pages 444–455. ACM Press, 2016.
2. M. Bellare, T. Ristenpart, P. Rogaway, and T. Stegers. Format-preserving encryption. In M. J. Jacobson Jr., V. Rijmen, and R. Safavi-Naini, editors, *SAC 2009*, volume 5867 of *LNCS*, pages 295–312. Springer, Heidelberg, Aug. 2009.
3. M. Bellare, P. Rogaway, and T. Spies. The FFX mode of operation for format-preserving encryption. Submission to NIST, Feb. 2010. `http://csrc.nist.gov/groups/ST/toolkit/BCM/documents/proposedmodes/ffx/ffx-spec.pdf`.
4. J. Black and P. Rogaway. Ciphers with arbitrary finite domains. In B. Preneel, editor, *CT-RSA 2002*, volume 2271 of *LNCS*, pages 114–130. Springer, Heidelberg, Feb. 2002.
5. E. Brier, T. Peyrin, and J. Stern. BPS: a format-preserving encryption proposal. Submission to NIST, 2010.
6. M. Brightwell and H. Smith. Using datatype-preserving encryption to enhance data warehouse security. In *20th National Information Systems Security Conference Proceedings (NISSC)*, pages 141–149, 1997.
7. S. Dara and S. Fluhrer. FNR: Arbitrary length small domain block cipher proposal. In *International Conference on Security, Privacy, and Applied Cryptography Engineering*, pages 146–154. Springer, 2014.
8. F. B. Durak and S. Vaudenay. Breaking and repairing the FF3 format preserving encryption over small domain. In *CRYPTO 2017*, pages 679–707. Springer, 2017.
9. M. Dworkin. Recommendation for Block Cipher Modes of Operation: Methods for Format-Preserving Encryption. *NIST Special Publication 800-38G*, Mar. 2016. `http://dx.doi.org/10.6028/NIST.SP.800-38G`.
10. M. Dworkin and R. Perlner. Analysis of VAES3 (FF2). Cryptology ePrint Archive, Report 2015/306, 2015. `http://eprint.iacr.org/2015/306`.
11. V. T. Hoang, B. Morris, and P. Rogaway. An enciphering scheme based on a card shuffle. In R. Safavi-Naini and R. Canetti, editors, *CRYPTO 2012*, volume 7417 of *LNCS*, pages 1–13. Springer, Heidelberg, Aug. 2012.
12. U. Mattsson. Format controlling encryption using datatype preserving encryption. Cryptology ePrint Archive, Report 2009/257, 2009. `http://eprint.iacr.org/2009/257`.
13. B. Morris and P. Rogaway. Sometimes-recurse shuffle - almost-random permutations in logarithmic expected time. In *EUROCRYPT 2014*, volume 8441 of *LNCS*. Springer, Heidelberg, May 2014.
14. R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.
15. M. Naor and O. Reingold. On the construction of pseudorandom permutations: Luby-Rackoff revisited. *Journal of Cryptology*, 12(1):29–66, 1999.
16. T. Ristenpart and S. Yilek. The mix-and-cut shuffle: Small-domain encryption secure against N queries. In R. Canetti and J. A. Garay, editors, *CRYPTO 2013, Part I*, volume 8042 of *LNCS*, pages 392–409. Springer, Heidelberg, Aug. 2013.
17. J. Vance. VAES3 scheme for FFX: An addendum to The FFX mode of operation for Format Preserving Encryption. Submission to NIST, May 2011.

A Proof of Lemma 5

For each $i \le t$, let X_i and X'_i be the round-i intermediate outputs of X and X' respectively. Let F_i be the round function at round i, and let $G_i(\cdot,\cdot) = F_i(K,\cdot,\cdot)$. For $Z \in \mathbb{Z}_M$, let $\mathrm{Match}_t(Z)$ be the event that $\mathbf{Left}(X_t) \boxminus \mathbf{Left}(X'_t) = Z$. We now give an induction proof on t that $\Pr[\mathrm{Match}_t(Z)] = 1/M$; the claimed result in the theorem statement is the special case $t = r$. First consider the base case $t = 2$. Note that

$$\begin{aligned}\mathbf{Left}(X_2) = \mathbf{Left}(X_1) = \mathbf{Left}(X) \boxplus G_1(T, \mathbf{Right}(X)), \text{ and}\\ \mathbf{Left}(X'_2) = \mathbf{Left}(X'_1) = \mathbf{Left}(X') \boxplus G_1(T, \mathbf{Right}(X'))\ .\end{aligned}$$

Since $\mathbf{Right}(X) \neq \mathbf{Right}(X')$, the random variables $\mathbf{Left}(X_2)$ and $\mathbf{Left}(X'_2)$ are independent and uniformly distributed over $\mathbb{Z}_M$. Hence $\Pr[\mathrm{Match}_2(Z)] = 1/M$.

Next, suppose that the claim above holds for $2, 4, \ldots, t-2$. We now prove that it holds for t. We consider the following two cases.

Case 1: $\mathbf{Left}(X_{t-3}) = \mathbf{Left}(X'_{t-3})$. Since X and X' are distinct and Feistel is a permutation, $X_{t-3} \neq X'_{t-3}$, and thus $\mathbf{Right}(X_{t-3}) \neq \mathbf{Right}(X'_{t-3})$. Since t is even,

$$\begin{aligned}\mathbf{Right}(X_{t-2}) &= \mathbf{Left}(X_{i-3}) \boxplus G_{t-2}(T, \mathbf{Right}(X_{t-3}))\\ &\neq \mathbf{Left}(X'_{i-3}) \boxplus G_{t-2}(T, \mathbf{Right}(X'_{t-3})) = \mathbf{Right}(X'_{t-2}).\end{aligned}$$

On the other hand, since t is even,

$$\begin{aligned}\mathbf{Left}(X_t) = \mathbf{Left}(X_{t-1}) = \mathbf{Left}(X_{t-2}) \boxplus G_{t-1}(T, \mathbf{Right}(X_{t-2})), \text{ and}\\ \mathbf{Left}(X'_t) = \mathbf{Left}(X'_{t-1}) = \mathbf{Left}(X'_{t-2}) \boxplus G_{t-1}(T, \mathbf{Right}(X'_{t-2}))\ .\end{aligned}$$

Since $\mathbf{Right}(X_{t-2}) \neq \mathbf{Right}(X'_{t-2})$, and G_{t-1} is a truly random function, independent of X_{t-2} and X'_{t-2}, the random variables $\mathbf{Left}(X_t)$ and $\mathbf{Left}(X'_t)$ are independently and uniformly distributed over $\mathbb{Z}_M$. Consequently $\mathbf{Left}(X_t) \boxminus \mathbf{Left}(X'_t)$ is uniformly distributed over $\mathbb{Z}_M$, and the probability that $\mathrm{Match}_t(Z)$ happens is $1/M$.

Case 2: $\mathbf{Left}(X_{t-3}) \neq \mathbf{Left}(X'_{t-3})$. Define Coll to be the event $\mathbf{Right}(X_{t-2}) = \mathbf{Right}(X'_{t-2})$. If Coll does not happen then using the same argument as in Case 1, we can show that the conditional probability that $\mathrm{Match}_t(Z)$ happens is $1/M$. Hence

$$\Pr[\mathrm{Match}_t(Z) \wedge \overline{\mathrm{Coll}}] = \frac{1}{M} \Pr[\overline{\mathrm{Coll}}]\ .$$

Consider the case that Coll does happen. Recall that

$$\begin{aligned}\mathbf{Left}(X_t) = \mathbf{Left}(X_{t-1}) = \mathbf{Left}(X_{t-2}) \boxplus G_{t-1}(T, \mathbf{Right}(X_{t-2})), \text{ and}\\ \mathbf{Left}(X'_t) = \mathbf{Left}(X'_{t-1}) = \mathbf{Left}(X'_{t-2}) \boxplus G_{t-1}(T, \mathbf{Right}(X'_{t-2}))\ .\end{aligned}$$

Subtracting side by side, we have

$$\mathbf{Left}(X_t) \boxminus \mathbf{Left}(X'_t) = \mathbf{Left}(X_{t-2}) \boxminus \mathbf{Left}(X'_{t-2})\ .$$

Hence

$$\Pr[\mathrm{Match}_t(Z) \wedge \mathrm{Coll}] = \Pr[\mathrm{Match}_{t-2}(Z) \wedge \mathrm{Coll}] \ .$$

We claim that the right-hand side is $\Pr[\mathrm{Coll}] \cdot \Pr[\mathrm{Match}_{t-2}(Z)]$. Summing up,

$$\begin{aligned}\Pr[\mathrm{Match}_t(Z)] &= \frac{1}{M}(1 - \Pr[\mathrm{Coll}]) + \Pr[\mathrm{Coll}] \cdot \Pr[\mathrm{Match}_{t-2}(Z)] \\ &= \frac{1}{M} + \Pr[\mathrm{Coll}] \cdot \left(\Pr[\mathrm{Match}_{t-2}(Z)] - \frac{1}{M}\right) = \frac{1}{M} \ .\end{aligned}$$

To justify the claim above, we need only prove that Coll and $\mathrm{Match}_{t-2}(Z)$ are conditionally independent, given $\mathbf{Left}(X_{t-3}) \neq \mathbf{Left}(X'_{t-3})$. Note that event $\mathrm{Match}_{t-2}(Z)$ depends solely on the round functions up to round $t-3$. Thus it suffices to show that regardless of the choices of the round functions from round 1 to round $t-3$, as long as $\mathbf{Left}(X_{t-3}) \neq \mathbf{Left}(X'_{t-3})$ then $\Pr[\mathrm{Coll}] = 1/N$. Recall that

$$\begin{aligned}\mathbf{Right}(X_{t-2}) &= \mathbf{Right}(X_{t-3}) \boxplus G_{t-2}(T, \mathbf{Left}(X_{t-3})), \text{ and} \\ \mathbf{Right}(X'_{t-2}) &= \mathbf{Right}(X'_{t-3}) \boxplus G_{t-2}(T, \mathbf{Left}(X'_{t-3})) \ .\end{aligned}$$

If $\mathbf{Left}(X_{t-3}) \neq \mathbf{Left}(X'_{t-3})$ then $\mathbf{Right}(X_{t-2})$ and $\mathbf{Right}(X'_{t-2})$ are independent and uniformly distributed over $\mathbb{Z}_N$. Hence regardless of the choices of the round functions from round 1 to round $t-3$, as long as $\mathbf{Left}(X_{t-3}) \neq \mathbf{Left}(X'_{t-3})$ then $\Pr[\mathrm{Coll}] = 1/N$.

B Proof of Lemma 7

Let F_i be the round function at round i, and let $G_i(\cdot,\cdot) = F_i(K,\cdot,\cdot)$. For any strings $Z \in \{0,1\}^{m+n}$ and $V \in \{0,1\}^m$ and for any integer $k \geq 1$, define $\mathsf{Diff}_k(Z)$ to be the event that $X_k \oplus X'_k = Z$, and let $\mathrm{Hit}_k(V)$ be the event that $\mathbf{Left}(X_k) \oplus \mathbf{Left}(X'_k) = V$. Fix an odd integer $t \geq 3$. We claim that for any string $V \in \{0,1\}^m \backslash \{0^m\}$,

$$\Pr[\mathsf{Diff}_t(V \,\|\, 0^n)] = \frac{\Pr[\mathrm{Hit}_{t-1}(V)]}{2^n} \ . \tag{5}$$

Moreover, for any strings Z and Z in $\{0,1\}^{m+n}$ such that $\mathbf{Right}(Z) \neq 0^n$ and $\mathbf{Right}(Z') \neq 0^n$, we claim that

$$\Pr[\mathsf{Diff}_t(Z)] - \Pr[\mathsf{Diff}_t(Z')] = \frac{\Pr[\mathsf{Diff}_1(U)] - \Pr[\mathsf{Diff}_1(U')]}{2^{(t-1)m/2}} \ , \tag{6}$$

where $U = 0^m \,\|\, \mathbf{Right}(Z)$ and $U' = 0^m \,\|\, \mathbf{Right}(Z')$. These claims will be justified later. We now use these to prove our results.

Case 1: $\mathbf{Right}(X) \neq \mathbf{Right}(X')$. For any odd integer $t \geq 3$ and for any $V \in \{0,1\}^m \backslash \{0^m\}$, from Lemma 5,

$$\Pr[\mathrm{Hit}_{t-1}(V)] = \frac{1}{2^m} \ .$$

Combining this with Equation (5), we have

$$\Pr[\mathsf{Diff}_t(V \,\|\, 0^n)] = \frac{1}{2^{m+n}} \ .$$

Next, let $R_0 = \mathbf{Right}(X) \oplus \mathbf{Right}(X')$. For any $R \in \{0,1\}^n \backslash \{0^n\}$, note that $\Pr[\mathsf{Diff}_1(0^m \,\|\, R)] = 0$ if $R \neq R_0$, and $\Pr[\mathsf{Diff}_1(0^m \,\|\, R)] = 1/2^m$ otherwise. Combining that with Equation (6), for any odd number $t \geq 3$, there must be a constant a_t such that $\Pr[\mathsf{Diff}_t(Z)] = a_t$ if $\mathbf{Right}(Z) \notin \{0^n, R_0\}$, and $\Pr[\mathsf{Diff}_t(Z)] = a_t + \frac{1}{2^{(t+1)m/2}}$ if $\mathbf{Right}(Z) = R_0$. To find this constant a_t, note that on the one hand,

$$\sum_{Z:\mathbf{Right}(Z)\neq 0^n} \Pr[\mathsf{Diff}_t(Z)] = (2^n - 1)2^m a_t + \frac{1}{2^{(t+1)m/2}} \ .$$

On the other hand,

$$\sum_{Z:\mathbf{Right}(Z)\neq 0^n} \Pr[\mathsf{Diff}_t(Z)] = 1 - \sum_{V\in\{0,1\}^m} \Pr[\mathsf{Diff}_t(V \,\|\, 0^n)] = 1 - \frac{2^m - 1}{2^{m+n}} \ .$$

Hence

$$a_t = \frac{1}{(2^n - 1)2^m}\Big(1 - \frac{2^m - 1}{2^{m+n}} - \frac{1}{2^{(t+1)m/2}}\Big) \ .$$

Thus for any odd $t \geq 7$,

$$a_t \geq \frac{1}{(2^n - 1)2^m}\Big(1 - \frac{2^m - 1}{2^{m+n}} - \frac{1}{2^{4m}}\Big) \geq \frac{1}{2^{m+n}} + \frac{1}{2^{2(m+n)}} \ ,$$

where the last inequality is due to the hypothesis that $m \geq \max\{2, n-1\}$. Hence for any $Z \in \{0,1\}^{m+n}$ such that $\mathbf{Right}(Z) \neq 0^n$ and any odd $t \geq 7$,

$$\Pr[\mathsf{Diff}_t(Z)] \geq \frac{1}{2^{m+n}} + \frac{1}{2^{2(m+n)}} \ .$$

Case 2: $\mathbf{Right}(X) = \mathbf{Right}(X')$. We need the following result from [1].

Lemma 9 ([1]). *Suppose that* $\mathbf{Right}(X) = \mathbf{Right}(X')$. *Then for any odd* $t \geq 3$ *and for any* $V, V' \in \{0,1\}^m \backslash \{0^m, L_0\}$, *where* $L_0 = \mathbf{Left}(X) \oplus \mathbf{Left}(X')$,

$$\Pr[\mathrm{Hit}_{t-1}(V)] = \Pr[\mathrm{Hit}_{t-1}(V')] \ .$$

Moreover,

$$\Pr[\mathrm{Hit}_{t-1}(L_0)] = \Pr[\mathrm{Hit}_{t-1}(V)] + \frac{1}{2^{(t-3)n/2}}, \ \textit{and}$$
$$\frac{2^n - 1}{2^{m+n} - 1} - \frac{1}{2^m \cdot 2^{(t-3)(m+n)/2}} \leq \Pr[\mathrm{Hit}_{t-1}(0^n)] \leq \frac{2^n - 1}{2^{m+n} - 1} \ .$$

For any $R \in \{0,1\}^n \backslash \{0^n\}$, note that $\Pr[\mathsf{Diff}_1(0^m \,\|\, R)] = 0$. Combining this with Equation (6), this means that there is a constant b_t such that $\Pr[\mathsf{Diff}_t(Z)] = b_t$, for every $Z \in \{0,1\}^{m+n}$ such that $\mathbf{Right}(Z) \neq 0^n$. Note that on the one hand,

$$\begin{aligned}\sum_{Z:\mathbf{Right}(Z)\neq 0^n} \Pr[\mathsf{Diff}_t(Z)] &= 1 - \sum_{V\in\{0,1\}^m\backslash\{0^m\}} \Pr[\mathsf{Diff}_t(V \,\|\, 0^n)] \\ &= 1 - \sum_{V\in\{0,1\}^m\backslash\{0^m\}} \frac{\Pr[\mathrm{Hit}_{t-1}(V)]}{2^n} \\ &= 1 - \frac{1 - \mathrm{Hit}_{t-1}(0^m)}{2^n} \ .\end{aligned}$$

On the other hand,

$$\sum_{Z:\mathbf{Right}(Z)\neq 0^n} \Pr[\mathsf{Diff}_t(Z)] = 2^m(2^n-1)b_t \ .$$

Using the bounds for $\mathrm{Hit}_{t-1}(0^m)$ as given in Lemma 9, we have

$$\frac{1}{2^{m+n}-1} \leq b_t \leq \frac{1}{2^{m+n}-1} + \frac{1}{(2^n-1)2^m \cdot 2^{(t-1)(m+n)/2}} \ ,$$

Hence for any $Z \in \{0,1\}^{m+n}$ such that $\mathbf{Right}(Z) \neq 0^n$,

$$\Pr[\mathsf{Diff}_t(Z)] = b_t \leq \frac{1}{2^{m+n}-1} + \frac{1}{(2^n-1)2^m \cdot 2^{(t-1)(m+n)/2}} \ .$$

Moreover,

$$\Pr[\mathsf{Diff}_t(Z)] = b_t \geq \frac{1}{2^{m+n}-1} \geq \frac{1}{2^{m+n}} + \frac{1}{2^{2(m+n)}} \ .$$

Next, to estimate $\Pr[\mathsf{Diff}_t(V \,\|\, 0^n)]$, from Lemma 9, we will need to establish both lower and upper bounds for $\mathrm{Hit}_{t-1}(V)$, for every $V \in \{0,1\}^m \backslash \{0^m\}$. Let $L_0 = \mathbf{Left}(X) \oplus \mathbf{Left}(X')$. Note that from Lemma 9, there is a constant c_t such that $\Pr[\mathrm{Hit}_{t-1}(V)] = c_t$ for every $V \in \{0,1\}^m \backslash \{0^m, L_0\}$. On the one hand,

$$\sum_{V\in\{0,1\}^m\backslash\{0^m\}} \Pr[\mathrm{Hit}_{t-1}(V)] = 1 - \Pr[\mathrm{Hit}_{t-1}(0^m)] \ .$$

On the other hand, by Equation (6),

$$\sum_{V\in\{0,1\}^m\backslash\{0^m\}} \Pr[\mathrm{Hit}_{t-1}(V)] = (2^m-1)c_t + \frac{1}{2^{(t-1)m/2}} \ .$$

Hence

$$c_t = \frac{1}{2^m-1}\Big(1 - \frac{1}{2^{(t-1)m/2}} - \Pr[\mathrm{Hit}_{t-1}(0^m)]\Big) \ .$$

Using the bounds of $\Pr[\mathrm{Hit}_{t-1}(0^n)]$ in Lemma 9 gives us

$$c_t \geq \frac{2^n}{2^{m+n}-1} - \frac{1}{(2^m-1)2^{(t-1)m/2}}, \text{ and}$$
$$c_t \leq \frac{2^n}{2^{m+n}-1} + \frac{1}{2^m(2^m-1)2^{(t-3)(m+n)/2}} \ ,$$

Combining this with Equation (5) we have

$$\Pr[\mathsf{Diff}_t(V \,\|\, 0^n)] \geq \frac{1}{2^{m+n}-1} - \frac{1}{2^n(2^m-1)2^{(t-1)m/2}}, \text{ and}$$
$$\Pr[\mathsf{Diff}_t(V \,\|\, 0^n)] \leq \frac{1}{2^{m+n}-1} + \frac{1}{(2^m-1)2^{(t-1)(m+n)/2}} \ .$$

For $t \geq 7$ and $m \geq \max\{2, n-1\}$,

$$\frac{1}{2^{m+n}-1} - \frac{1}{2^n(2^m-1)2^{(t-1)m/2}} \geq \frac{1}{2^{m+n}} + \frac{1}{2 \cdot 2^{2(m+n)}} \ ,$$

and thus

$$\Pr[\mathsf{Diff}_t(V \,\|\, 0^n)] \geq \frac{1}{2^{m+n}} + \frac{1}{2 \cdot 2^{2(m+n)}} \ .$$

On the other hand,

$$\Pr[\mathrm{Hit}_{t-1}(L_0)] = c_t + \frac{1}{2^{(t-1)m/2}} \geq \frac{2^n}{2^{m+n}-1} + \frac{1 - 1/(2^m-1)}{2^{(t-1)m/2}} \ .$$

Combining this with Equation (5) we have

$$\Pr[\mathsf{Diff}_t(L_0 \,\|\, 0^n)] \geq \frac{1}{2^{m+n}-1} + \frac{1 - 1/(2^m-1)}{2^n \cdot 2^{(t-1)m/2}} \ .$$

Justifying Equation (5). Fix an odd integer $t \geq 3$ and a string V in the set $\{0,1\}^n \backslash \{0^n\}$. Since t is odd, $\mathbf{Right}(X_t)$ is also $\mathbf{Right}(X_{t-1})$, and the same holds for X'. Note that $\mathsf{Diff}_t(V \,\|\, 0^n)$ happens if and only if $\mathbf{Right}(X_{t-1}) = \mathbf{Right}(X'_{t-1})$ and $\mathrm{Hit}_{t-2}(V)$ happens. On the other hand, if $\mathrm{Hit}_{t-2}(V)$ happens then $\mathbf{Left}(X_{t-2}) \neq \mathbf{Left}(X_{t-2})$, and thus $\mathbf{Right}(X_{t-1})$ and $\mathbf{Right}(X'_{t-1})$ are independent random strings. Hence

$$\Pr[\mathsf{Diff}_t(V \,\|\, 0^n)] = \frac{\Pr[\mathrm{Hit}_{t-2}(V)]}{2^n} \ .$$

Justify Equation (6). Fix an odd integer $t \geq 3$ and fix a string $Z \in \{0,1\}^{m+n}$ such that $\mathbf{Right}(Z) \neq 0^n$. First consider the case that $\mathrm{Hit}_{t-2}(0^m)$ does not happen. Then $\mathsf{Diff}_t(Z)$ happens iff $\mathbf{Right}(X_{t-1}) = \mathbf{Right}(X'_{t-1}) \oplus \mathbf{Right}(Z)$ and $\mathrm{Hit}_t(\mathbf{Left}(Z))$ happens. If $\mathrm{Hit}_{t-2}(0^m)$ does not happen then $\mathbf{Right}(X_{t-1})$ and $\mathbf{Right}(X'_{t-1})$ are independent random strings, and thus

$$\Pr[\mathbf{Right}(X_{t-1}) = \mathbf{Right}(X'_{t-1}) \oplus \mathbf{Right}(Z) \mid \neg \mathrm{Hit}_{t-2}(0^m)] = 2^{-n} \ .$$

Given that $\mathbf{Right}(X_{t-1}) = \mathbf{Right}(X'_{t-1}) \oplus \mathbf{Right}(Z)$ and $\mathrm{Hit}_{t-2}(0^m)$ does not happen, the strings $\mathbf{Left}(X_t)$ and $\mathbf{Left}(X'_t)$ are conditionally independent random strings, and thus $\mathrm{Hit}_t(\mathbf{Left}(Z))$ occurs with conditional probability 2^{-m}. Hence

$$\Pr[\mathsf{Diff}_t(Z) \mid \neg\mathrm{Hit}_{t-2}(0^m)] = 2^{-(m+n)} \ .$$

Now consider the case that $\mathrm{Hit}_{t-2}(0^m)$ happens. Then $\mathsf{Diff}_t(Z)$ happens if and only if $\mathsf{Diff}_{t-2}(U)$ and $\mathrm{Hit}_t(\mathbf{Left}(Z))$ happen, where $U = 0^m \,\|\, \mathbf{Right}(Z)$. Note that $\mathsf{Diff}_{t-2}(U)$ implies that $\mathbf{Right}(X_{t-1}) \neq \mathbf{Right}(X'_{t-1})$ and $\mathrm{Hit}_{t-2}(0^m)$ happens. Hence given that the event $\mathsf{Diff}_{t-2}(U)$ happens, $\mathrm{Hit}_t(\mathbf{Left}(Z))$ happens with conditional probability 2^{-m}, and thus

$$\Pr[\mathsf{Diff}_t(Z) \wedge \mathrm{Hit}_{t-2}(0^m)] = \frac{\Pr[\mathsf{Diff}_{t-2}(U)]}{2^m} \ .$$

Summing up,

$$\Pr[\mathsf{Diff}_t(Z)] = \frac{\Pr[\mathsf{Diff}_{t-2}(U)]}{2^m} + \frac{\Pr[\neg\mathrm{Hit}_{t-2}(0^m)]}{2^{m+n}} \ .$$

By repeating the argument above, we will be able to express $\Pr[\mathsf{Diff}_{t-2}(U)]$ via $\Pr[\mathsf{Diff}_{t-4}(U)]$ and $\Pr[\neg\mathrm{Hit}_{t-4}(0^n)]$ and so on. Hence there is a constant d_t such that

$$\Pr[\mathsf{Diff}_t(Z)] = d_t + \frac{\Pr[\mathsf{Diff}_1(U)]}{2^{(t-1)m/2}} \ .$$

Thus for any Z and Z' in $\{0,1\}^{m+n}$ such that $\mathbf{Right}(Z) \neq 0^n$ and $\mathbf{Right}(Z') \neq 0^n$, we have

$$\Pr[\mathsf{Diff}_t(Z)] - \Pr[\mathsf{Diff}_t(Z')] = \frac{\Pr[\mathsf{Diff}_1(U)] - \Pr[\mathsf{Diff}_1(U')]}{2^{(t-1)m/2}},$$

where $U' = 0^m \,\|\, \mathbf{Right}(Z')$. This completes the proof.

C Proof of Theorem 8

First we show that $\mathbf{Adv}^{\mathsf{mg}}_{\mathsf{XS}} \leq 1/d^m$. Consider an arbitrary simulator $\mathcal{S}$. To win the game, $\mathcal{S}$ must find the target Z but is given only the tweaks. As Z is uniformly distributed in $(\mathbb{Z}_d)^m$ independent of the tweaks, the chance that the simulator can find Z is at most $1/d^m$. Next, we show that

$$\begin{aligned}\Pr[\mathbf{G}^{\mathsf{mr}}_{\mathsf{F},\mathsf{XS}}(\mathsf{DD})] \geq 1 - \frac{(q \cdot \lceil m/r \rceil)^2}{2 \cdot D^{n-1}} - ms \cdot \exp\Big(\frac{-q(d-s)^2}{2Dd(D+d-s)}\Big) \\ -m(d-s) \cdot \exp\Big(\frac{-qs^2}{3Dd(D+s)}\Big) \ . \qquad (7)\end{aligned}$$

Assume that the $q \cdot \lceil m/r \rceil$ inputs to F are distinct, so that the outputs of F are independently and uniformly distributed over $(\mathbb{Z}_D)^n$. This happens with

probability at most $\frac{(q\cdot\lceil m/r\rceil)^2}{2\cdot D^{n-1}}$. Now, it suffices to show that for any digit of $Z = z_1 \cdots z_m$, the attack recovers it with probability at least

$$1 - s \cdot \exp\Big(\frac{-q(d-s)^2}{2Dd(D+d-s)}\Big) - (d-s)\cdot\exp\Big(\frac{-qs^2}{3Dd(D-s)}\Big), \tag{8}$$

since Equation (7) above then follows by a union bound. Without loss of generality, consider the first digit of Z. Let $V_{i,a}$ be the Bernoulli random variable that the first digit of the i-th ciphertext is a. Now, in the frequency histogram of the first digit of the ciphertexts, the height of the column corresponding to the value a is $V_{1,a} + \cdots + V_{q,a}$. We claim that

(i) For any number $a \in S = \{z_1 \bmod d, z_1 + 1 \bmod d, \ldots, z_1 + (D \bmod d) - 1 \bmod d\}$,

$$\Pr[V_{1,a} + \cdots + V_{q,a} \le q/d] \le \exp\Big(\frac{-q(d-s)^2}{2Dd(D+d-s)}\Big)\ .$$

In other words, a supposedly tall column in the frequency histogram is unlikely to appear short; the height of this column is likely to be bigger than the average.

(ii) For $a \in \mathbb{Z}_d \backslash S$,

$$\Pr[V_{1,a} + \cdots + V_{q,a} \ge q/d] \le \exp\Big(\frac{-qs^2}{3Dd(D-s)}\Big)\ .$$

In other words, a supposedly short column in the empirical histogram is unlikely to appear tall; the height of this column is likely to be smaller than the average.

Equation (8) then follows (i) and (ii) by a union bound. We now justify the claims (i) and (ii) above. First consider claim (i). Fix $a \in S$. Let $\mu = \frac{\lceil D/d\rceil}{D}$ and $\epsilon = \frac{d-s}{d\cdot\lceil D/d\rceil}$. Note that for any $i \le q$, we have $V_{i,a} = 1$ with probability μ, and $q/d = (1-\epsilon)q\mu$. Since $V_{1,a}, \ldots, V_{q,a}$ are independent and identically distributed Bernoulli random variables, by using Chernoff's bound,

$$\Pr[V_{1,a} + \cdots + V_{q,a} \le (1-\epsilon)q\mu] \le \exp\Big(\frac{-\epsilon^2 q\mu}{2}\Big) = \exp\Big(\frac{-q(d-s)^2}{2Dd(D+d-s)}\Big)\ .$$

Next, consider claim (ii). Fix $a \in \mathbb{Z}_d \backslash S$. Let $\mu^* = \frac{\lfloor D/d\rfloor}{D}$ and $\epsilon^* = \frac{s}{d\cdot\lfloor D/d\rfloor}$. Note that for any $i \le q$, we have $V_{i,a} = 1$ with probability μ^*, and $q/d = (1+\epsilon^*)q\mu^*$. Since $V_{1,a}, \ldots, V_{q,a}$ are independent and identically distributed Bernoulli random variables, by using Chernoff's bound,

$$\Pr[V_{1,a} + \cdots + V_{q,a} \ge (1+\epsilon^*)q\mu^*] \le \exp\Big(\frac{-(\epsilon^*)^2 q\mu^*}{3}\Big) = \exp\Big(\frac{-qs^2}{3Dd(D-s)}\Big)\ .$$

This completes the proof.