Generic Attacks for the Xor of k Random Permutations

Jacques Patarin

Université de Versailles
45 avenue des Etats-Unis
78035 Versailles Cedex - France

Abstract. Xoring the output of k permutations, $k \geq 2$ is a very simple way to construct pseudo-random functions (PRF) from pseudo-random permutations (PRP). Moreover such construction has many applications in cryptography (see [2–5] for example). Therefore it is interesting both from a theoretical and from a practical point of view, to get precise security results for this construction. In this paper, we will describe the best attacks that we have found on the Xor of k random n-bit to n-bit permutations. When $k = 2$, we will get an attack of computational complexity $O(2^n)$. This result was already stated in [2]. On the contrary, for $k \geq 3$, our analysis is new. We will see that the best known attacks require much more than 2^n computations when not all of the 2^n outputs are given, or when the function is changed on a few points. We obtain like this a new and very simple design that can be very useful when a security larger than 2^n is wanted, for example when n is very small.

Key words: Pseudorandom functions, pseudorandom permutations, Luby-Rackoff backwards, generic attacks

1 Introduction

The problem of converting pseudorandom permutations (PRP) into pseudorandom functions (PRF) named "Luby-Rackoff backwards" was first considered in [3]. This problem is obvious if we are interested in an assymptotical security model (since a PRP is then a PRF), but not if we are interested in achieving more optimal and concrete security bounds. More precisely, the loss of security when regarding a PRP as a PRF comes from the "birthday attack" which can distinguish a random permutation from a random function of n bits to n bits, in $2^{\frac{n}{2}}$ operations and $2^{\frac{n}{2}}$ queries. In [5] (Theorem 2 p.474), it has been proved that the Xor of k PRP gives a PRF with security at least in $O(2^{\frac{k}{k+1}n})$. (For $k = 2$ this gives $O(2^{\frac{2}{3}n})$). Moreover in [2], it has been proved that the Xor of two PRP gives a PRF with security at least in $O(2^n/n^{\frac{2}{3}})$ and at most in $O(2^n)$, which is much better than the birthday bound in $O(2^{\frac{n}{2}})$. Similarly in [8], it has been proved that in fact the security is at least in (and therefore exactly in) $O(2^n)$ for this problem to distinguish the Xor of two PRP from a PRF. An

interesting question is "Can we hope to get even better bound than $O(2^n)$ with more than two Xor, particularly if not all the 2^n inputs/outputs are given to the cryptanalysis ?" In this paper, we will study this question. Let F_k denote the Xor of k random permutations. Let G_k denote the function F_k except on a few secret (or public) points x_i where $G(x_i)$ is random (for example it can be only the point 0). We will distinguish 4 kinds of attack scenarios:

1. The adversary has access to the full codebook of F_k, i.e. exactly all the 2^n pairs of function input and function output.

2. The adversary has access to almost, but not all, the entire codebook of F_k, i.e. to m pairs with $m \simeq 2^n$ and $m < 2^n$.

3. The adversary wants to attack G_k (instead of F_k) and he has access to the full codebook of G_k.
   Moreover, in these scenarios 2 and 3 we will also assume that the adversary has access to a generator of such functions F_k (or G_k), i.e. has access to μ such functions and he wants to distinguish these μ functions from μ random independent functions.

4. Finally, in scenario 4, we will be as in scenario 2 except that:
   a. The adversary has access to only one function F_k (not a generator).
   b. We look in this scenario 4 for the best Advantage that the adversary can get even if this mathematical value is $\ll 1$ (and therefore cannot be used to distinguish).

To analyze these scenarios, we will introduce what we call "stable" attacks and "unstable" attacks. An attack will be called "stable" if the attack is still valid with a similar complexity when a few points of the functions are changed to truly random values. We will present the best "stable" and "unstable" attacks that we have found on the Xor of k functions, $k \geq 2$ when we study a generator of such functions (not only one such function). We will see that in Scenario 1, the best security bound is indeed in $O(2^n)$, but in Scenario 2 and 3, the best attacks have an even greater complexity. So it gives candidate schemes to build PRF from PRP in a still very simple way and with potentially even better security. Since building PRF from PRP has many applications (see [2–4]), we think that these results are really interesting both from theoretical and from practical point of view.

The paper is organised as follows. We will analyse Scenario 1 in section 2, Scenario 2 in section 3 and 4, Scenario 3 in section 5 and Scenario 4 in sections 6 and 7. Then we will analyse the case where the k Xor are done on only one permutation (instead of k independant permutations) in section 8. Some other variants and open problems are presented in section 9. Finally, the results obtained are summarized in section 10. We have decided to present in Appendices the computation of all the mean values and standard deviations needed.

2 Scenario 1 on $f_1 \oplus f_2 \oplus \ldots \oplus f_k$ with $O(2^n)$ computations

Notations: In all this paper we will denote $I_n = \{0,1\}^n$. F_n will be the set of all applications from I_n to I_n, and B_n the set of all permutations from I_n to I_n. So $|I_n| = 2^n$, $|F_n| = 2^{n \cdot 2^n}$, and $|B_n| = (2^n)!$. $x \in_R A$ will mean that x is randomly chosen in A, with a uniform distribution.

Aim: In this section we want to distinguish $f \oplus g$, with $f, g \in_R B_n$ from $h \in_R F_n$.

Attack. We analyze a function G, we want to know if $G = f \oplus g$, $f, g \in_R B_n$, or if $G = h$, $h \in_R F_n$. If we have access to all the 2^n values $G(x)$, then we can compute $T = \oplus_{i=1}^{2^n} G(i)$. If $G = f \oplus g$, then with probability 1, we have $T = 0$. (Proof: If f is a permutation we have $\oplus_{i=1}^{2^n} f(i) = \oplus_{i=1}^{2^n} i = 0$ and similarly $\oplus_{i=1}^{2^n} g(i) = 0$, so $\oplus_{i=1}^{2^n} f(i) \oplus g(i) = 0$). If $G = h$, $h \in_R F_n$, then we have $T = 0$ with probability $\frac{1}{2^n}$. Therefore, by computing T, we can distinguish $f \oplus g$ from h with a very good probability. This attack is in $O(2^n)$ computations, with $O(2^n)$ input/output values.

Aim with $k \geq 3$: We want to distinguish $f_1 \oplus f_2 \oplus \ldots \oplus f_k$, with $f_1, f_2, \ldots, f_k \in_R B_n$ from $h \in_R F_n$.
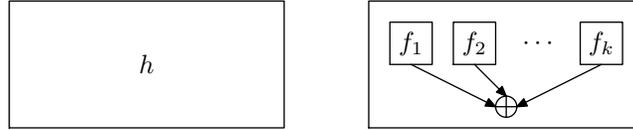

Fig. 1. Our attack distinguish between a function and the xor of k permutations

Attack. We use exactly the same attack: by computing T, we can distinguish $f_1 \oplus f_2 \oplus \ldots \oplus f_k$ from h with a very good probability. This attack is in $O(2^n)$ computations, with $O(2^n)$ input/output values.

Therefore, it seems that 2^n is the best security result that we can get with k Xor of permutations, for all k. However we can notice that if instead of having $f_1 \oplus f_2 \oplus \ldots \oplus f_k$, we use a function G such that $G = f_1 \oplus f_2 \oplus \ldots \oplus f_k$ except on a few points (or even except only on 0), and on these few points the output of G is truly random, then the above attack fails. We will say that this attack is "unstable". More precisely, we will define "stable" attacks as follows:

Definition We want to distinguish a function G of F_n (generated by a function generator) from truly random functions $f \in_R F_n$ with an attack A. Let $P(n)$ be a polynomial in n and $x_1, \ldots x_\phi$ be ϕ points randomly chosen in I_n with $\phi \leq P(n)$. Let $\Phi = \{x_1, \ldots x_\phi\}$. Let $G' = G$ on all the points of $F_n - \Phi$ and $G'(x_i)$ be truly random on all $x_i \in \Phi$. Then if for each such sets Φ the attack A is polynomial (in n) against G', we will say that this attack is stable on G.

Remark: It is possible to store a few random points with $O(n)$ random bits, i.e. polynomial in n, but to store a random function of F_n, we need $n \cdot 2^n$ random

bits, i.e. not polynomial in n. To avoid an "unstable" attack on G, we have to change the design of G only on a few points. However to avoid a "stable" attack on G, the design of G must be deeply changed.

3   Scenario 2 on $f \oplus g$ with $O(2^{2n})$ computations

Aim: we want to distinguish a generator A of functions $f \oplus g$, with $f, g \in_R B_n$, from a generator B of functions h, with $h \in_R F_n$; i.e. we can have access to more than one test function G, these G functions are generated from A or from B and we have to distinguish these two cases with a non negligible probability. Moreover for each G function, we have access to all the inputs/outputs, except a few points. (Or alternatively, from generator A, $G = f \oplus g$ except on a few points).

Attack. We will count the number N of collisions on the functions G. Therefore if we have access to m inputs/outputs for G, $G(x_i) = y_i$ for $1 \leq i \leq m$, N is the number of (i,j), $1 \leq i < j \leq m$ such that $G(x_i) = G(x_j)$. (In our attack we will generally choose $m \simeq 2^n$ but we will not need $m = 2^n$.)

Case of random functions. We know that for a random function of F_n, we have $E(N) = \frac{m(m-1)}{2 \cdot 2^n}$ and $\sigma(N) = O(\frac{m}{\sqrt{2^n}})$ where $E(N)$ denotes the mean value of N, and $\sigma(N)$ denotes the standard deviation of N. (See Appendix A for the proof of these results). Therefore, for a generator with μ such functions,

$$E(N) = \frac{\mu \cdot m(m-1)}{2 \cdot 2^n} \quad \text{and} \quad \sigma(N) = O(\frac{\sqrt{\mu} \cdot m}{\sqrt{2^n}})$$

(Since if $X_1, \ldots, X_n$ are n independent events with $E(X_i) = E$ and $\sigma(X_i) = \sigma$, we have $E(X_1 + \ldots + X_n) = nE$ and $\sigma(X_1 + \ldots X_n) = \sqrt{n}\sigma$. Here the generator generates independent functions $h_1, \ldots h_n$).

Case of $f \oplus g$. We know that if $G = f \oplus g$, with $f, g \in_R B_n$, we have $E(N) = \frac{m(m-1)}{2} \cdot \frac{1}{2^n - 1}$ and $\sigma(N) = O(\frac{m}{\sqrt{2^n}})$, (see Appendix B for the proof of these results). Therefore, for a generator with μ such functions,

$$E(N) = \frac{\mu \cdot m(m-1)}{2} \cdot \frac{1}{2^n - 1}$$

(This shows that we have in average slightly more collisions with $f \oplus g$ than with h), and

$$\sigma(N) = O(\frac{\sqrt{\mu}m}{\sqrt{2^n}})$$

From Bienayme-Tchebichev theorem we know that we will be able to distinguish h from $f \oplus g$ with a good probability when

$$\sigma(N)_h << |E(N)_h - E(N)_{f \oplus g}|$$

and

$$\sigma(N)_{f \oplus g} << |E(N)_h - E(N)_{f \oplus g}|$$

(This is a sufficient condition to distinguish h from $f \oplus g$.)

Here these conditions give:

$$\frac{\sqrt{\mu}m}{\sqrt{2^n}} << \frac{\mu \cdot m(m-1)}{2 \cdot 2^{2n}}$$

For $m \simeq 2^n$, this gives: $\mu \geq 2^n$ and the complexity of this attack is in $O(\mu \cdot m)$ computations, i.e. in $O(2^{2n})$.

Conclusion: This is a "stable attack" on $f \oplus g$ with $O(2^{2n})$ computations (see section 5 to see why this attack is "stable").

Remark: This is the best "stable" generic attack on $f \oplus g$ that we have found.

4   Scenario 2 on $f_1 \oplus f_2 \oplus \ldots \oplus f_k$ with $O(2^{(2k-2)n})$ computations

Aim: we want to distinguish a generator A of functions $f_1 \oplus f_2 \oplus \ldots \oplus f_k$, with $f_1, \ldots f_k \in_R B_n$ from a generator B of functions $h \in_R F_n$. We assume that we have access to m inputs/outputs values for each function G, with $m \neq 2^n$ (but $m \simeq 2^n$ if we want), i.e. we look for a stable attack (the attack will still be valid if a few inputs/outputs of G are changed).

Remark: Section 3 was a special case of section 4 with $k = 2$.

Attack. We will count the number N of collisions on all the functions G. Therefore, if we have access to m inputs/outputs for each function G, N is the number of (i, j), $i < j$, such that: $G(x_i) = G(x_j)$.

Case of random functions. We have seen in Section 3 (and in Appendix B) that for a random function of F_n, we have:

$$E(N) = \frac{m(m-1)}{2 \cdot 2^n} \quad \text{and} \quad \sigma(N) = O(\frac{m}{\sqrt{2^n}})$$

Therefore, for a generator with μ such functions,

$$E(N) = \frac{\mu \cdot m(m-1)}{2 \cdot 2^n} \quad \text{and} \quad \sigma(N) = O(\frac{\sqrt{\mu} \cdot m}{\sqrt{2^n}})$$

Case of $f_1 \oplus f_2 \oplus \ldots \oplus f_k$. We know that if $G = f_1 \oplus f_2 \oplus \ldots \oplus f_k$, with $f_1, f_2, \ldots f_k \in_R B_n$, we have

$$E(N) = \frac{m(m-1)}{2} \cdot \frac{1}{2^n}\Big[1 + \frac{(-1)^k}{(2^n - 1)^{k-1}}\Big]$$

and $\sigma(N) = O(\frac{m}{\sqrt{2^n}})$, (Proof: see Appendix C). Therefore, for a generator with μ such functions,

$$E(N) = \frac{\mu \cdot m(m-1)}{2} \cdot \frac{1}{2^n}\Big[1 + \frac{(-1)^k}{(2^n - 1)^{k-1}}\Big] \quad \text{and} \quad \sigma(N) = O(\frac{\sqrt{\mu}m}{\sqrt{2^n}})$$

From Bienayme-Tchebichev theorem we know that we will be able to distinguish h from $f_1 \oplus f_2 \oplus \ldots \oplus f_k$ with a good probability when

$$\sigma(N)_h << |E(N)_h - E(N)_{f_1 \oplus \ldots \oplus f_k}|$$

and

$$\sigma(N)_{f_1 \oplus \ldots \oplus f_k} << |E(N)_h - E(N)_{f_1 \oplus \ldots \oplus f_k}|$$

(This is a sufficient condition to distinguish h from $f_1 \oplus \ldots \oplus f_k$).

Here these conditions give:

$$\frac{\sqrt{\mu}m}{\sqrt{2^n}} << \frac{\mu \cdot m^2}{2^{kn}}$$

For $m \simeq 2^n$, this gives: $\mu \geq 2^{(2k-3)n}$ and therefore the complexity of this attack is in $O(\mu \cdot m)$ computations, i.e. in $O(2^{(2k-2)n})$.

5 Analysis of Scenario 3

Let G^* be perfectly random on φ points, and $G^*(x) = f_1(x) \oplus f_2(x) \oplus \ldots \oplus f_k(x)$, with $f_1, \ldots, f_k \in_R B_n$, on the $2^n - \varphi$ other points. Let ϕ be the set of the φ special points. Let assume that we know G^* on m points x_i, such that φ' of these point are in Φ and $m - \varphi'$ are not in Φ, $\varphi' \leq \varphi$. Let N be the number of collisions $G^*(x_i) = G^*(x_j)$, with $i < j$. We have: $N = N_1 + N_2 + N_3$ with

$N_1 =$ number of collisions with $x_i \notin \phi$ and $x_j \notin \phi$, $i < j$.
$N_2 =$ number of collisions with $x_i \notin \phi$ and $x_j \in \phi$, $i < j$.
$N_3 =$ number of collisions with $x_i \in \phi$ and $x_j \in \phi$, $i < j$.

We have $E(N) = E(N_1) + E(N_2) + E(N_3)$. From Theorem 1 of Appendix C, we have:

$$E(N_1) = \frac{(m - \varphi')(m - \varphi' - 1)}{2} \cdot \frac{1}{2^n} [1 + \frac{(-1)^k}{(2^n - 2)^{k-1}}]$$

Moreover, $E(N_2) = \frac{\varphi'(m - \varphi')}{2^n}$ and $E(N_3) = \frac{\varphi'(\varphi' - 1)}{2 \cdot 2^n}$. Therefore

$$E(N) = \frac{m(m-1)}{2} \cdot \frac{1}{2^n} + \frac{(m - \varphi')(m - \varphi' - 1)}{2} \cdot \frac{1}{2^n} \frac{(-1)^k}{(2^n - 1)^{k-1}}$$

So if $m \simeq 2^n$ and $\varphi << 2^n$, we have $\varphi' << 2^n$ and

$$|E(N)_{G^*} - E(N)_{f \in_R F_n}| \simeq \frac{1}{2 \cdot (2^n - 1)^{k-2}}$$

Therefore this attack by counting N for G^* will work with the same complexity as the attack by counting N on $f_1(x) \oplus f_2(x) \oplus \ldots \oplus f_k(x)$ as long as $\varphi << 2^n$, so we say that this attack is "stable". (This also means that "scenario 3" and "scenario 2" have the same conplexity).

6 Scenario 4: Best Known Advantage on a single $f \oplus g$ with $m < 2^n$

Let h be the single function of F_n that we want to study. h can be $h \in_R F_n$, or h can be $h = f \oplus g$ with $f, g \in_R B_n$. We assume that we know h on m points x_i: $h(x_i) = y_i$, $\forall i$, $1 \le i \le m$. Let N be the number of collisions on these m points, i.e. N is the number of (i, j), $1 \le i < j \le m$ such that: $y_i = y_j$.
First case: $m \ll \sqrt{2^n}$. Let ϕ be this attack:
• if $N = 0$ then ϕ outputs 0.
• if $N \ne 0$ (i.e. $N \ge 1$) then ϕ outputs 1.
Let $p_1 = Pr_{h \in_R F_n}(\phi(h) = 1)$, and $p_1^* = Pr_{f, g \in_R B_n}(\phi(f \oplus g) = 1)$. If $m \ll \sqrt{2^n}$, $p_1 \simeq \frac{m(m-1)}{2.2^n}$, and $p_1^* \simeq \frac{m(m-1)}{2.2^n}(1 + \frac{1}{2^n})$ (cf Appendix B). Therefore, if $m \ll \sqrt{2^n}$, $Adv(\phi) = |p_1 - p_1^*| \simeq \frac{m(m-1)}{2.2^{2n}}$. This shows that if $m \ll \sqrt{2^n}$, the Advantage, Adv_m to distinguish $h \in_R F_n$ from $f \oplus g$, $f, g \in_R B_n$ is at least in $O(\frac{m(m-1)}{2^{2n}})$. (This value is $\ll 1$ and therefore too small to distinguish).
Remark. When $m = 1$, $m = 2$ and $m = 3$, the exact values for Adv_m are given in [9]. More precisely in [9], it is shown that $Adv_1 = 0$, $Adv_2 = \frac{1}{2^n(2^n-1)} \simeq \frac{1}{2^{2n}}$, $Adv_3 = \frac{1}{2^{2n}}(\frac{3.2^{2n} - 12.2^n + 4}{(2^n-1)(2^n-2)}) \simeq \frac{3}{2^{2n}}$.
Second case: $\sqrt{2^n} \ll m \ll 2^n$. Let Ψ be this attack:
• if $N \ge \frac{m(m-1)}{2.2^n}$, then Ψ outputs 1.
• if $N < \frac{m(m-1)}{2.2^n}$, then Ψ outputs 0.
(Ψ is a "2-point" attack). If $(f, g) \in_R B_n$ we have $E(N) = \frac{m(m-1)}{2.(2^n-1)} \simeq \frac{m(m-1)}{2.2^n}(1 + \frac{1}{2^n})$ and $\sigma(N) \simeq \frac{m}{\sqrt{2}\sqrt{2^n}}$ (cf Appendix B). If $\sqrt{2^n} \ll m \ll 2^n$, then the distribution of N is similar to the Gaussian distribution of density $\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x - E(N))^2}{2\sigma^2}}$. Therefore we have: $Adv(\Psi) = O(\frac{\Delta E(N)}{\sigma})$, $Adv(\Psi) = O(\frac{m}{2^{\frac{3n}{2}}})$. This shows that if $\sqrt{2^n} \ll m \ll 2^n$, the Advantage to distinguish $h \in_R F_n$ from $f \oplus g$, $f, g \in_R B_n$ is at least $O(\frac{m}{2^{\frac{3n}{2}}})$. (This value is $\ll 1$, this is why in scenarios 2 and 3 we used a generator of functions).

7 Scenario 4: Best Known Advantage on $f_1 \oplus f_2 \oplus \ldots \oplus f_k$ with $m < 2^n$

First case: $m \ll \sqrt{2^n}$. Let ϕ be the attack ϕ seen in section 6. Let $p_1 = Pr_{h \in_R F_n}(\phi(h) = 1)$ as in section 6. let $p_1^*(k) = Pr_{f_1, \ldots, f_k \in_R B_n}(\phi(f_1 \oplus f_2 \oplus \ldots f_k) = 1)$. If $m \ll \sqrt{2^n}$, $p_1 \simeq \frac{m(m-1)}{2.2^n}$ and $p_1^*(k) \simeq \frac{m(m-1)}{2.2^n}(1 + \frac{(-1)^k}{(2^n-1)^{k-1}})$ (cf Appendix C). Therefore if $m \ll \sqrt{2^n}$, $Adv(\phi) = |p_1 - p_1^*(k)| \simeq \frac{m(m-1)}{2.2^{kn}}$. This shows that if $m \ll \sqrt{2^n}$, the Advantage, Adv_m to distinguish $h \in_R F_n$ from $f_1 \oplus f_2 \oplus \ldots \oplus f_k$, $f_1, \ldots, f_k \in_R B_n$, is at least $O(\frac{m(m-1)}{2^{kn}})$.
Second case: $\sqrt{2^n} \ll m \ll 2^n$. Let Ψ be the attack Ψ seen in section 6. If $h \in_R F_n$ we have $E(N) = \frac{m(m-1)}{2.2^n}$. If $f_1, \ldots, f_k \in_R B_n$ we have $E(N) = \frac{m(m-1)}{2.2^n}(1 +$

$\frac{(-1)^k}{(2^n-1)^{k-1}}$) and $\sigma(N) \simeq \frac{m}{\sqrt{2}\sqrt{2^n}}$ (cf Appendix C). If $\sqrt{2^n} \ll m \ll 2^n$, then the distribution of N is similar to the Gaussian distribution of density $\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-E(N))^2}{2\sigma^2}}$.
Therefore we have: $Adv(\Psi) = O(\frac{\Delta E(N)}{\sigma})$, $Adv(\Psi) = O(\frac{m}{2^{(k-\frac{1}{2})n}})$. This shows that if $\sqrt{2^n} \ll m \ll 2^n$, the Advantage to distinguish $h \in_R F_n$ from $f \oplus \ldots \oplus f_k$, $f, \ldots, f_k \in_R B_n$ is at least $O(\frac{m}{2^{(k-\frac{1}{2})n}})$. (This value is $\ll 1$, this is why in scenarios 2 and 3 we used a generator of functions).

8 A simple variant of the schemes with only one permutation

Variant with 2 Xor

Instead of $G = f_1 \oplus f_2$, $f_1, f_2 \in_R B_n$, we can study $G'(x) = f(x\|0) \oplus f(x\|1)$, with $f \in_R B_n$ and $x \in I_{n-1}$. This variant was already introduced in [2]. There are many common results between G and G' but also a few differences. It is possible to prove that our attacks (stable and unstable) on G are also valid on G' with similar properties. The (unstable) attack of Section 2 in $O(2^n)$ is also valid for G', since $\oplus_{x=1}^{2^n} G'(x) = \oplus_{i=1}^{2^n} i = 0$, and the number of collisions for the (stable) attacks of Section 3 will be similar for G and G'.

A specific attack on G'

There is however a specific attack on G' that do not exist on G since $\forall x \in I_n, G'(x) \neq 0$. Therefore, if we know m outputs y_i of G, we can test if $\forall i, 1 \leq i \leq m, y_i \neq 0$ (#). The probability of this event is 1 on G' and $(1 - \frac{1}{2^n})^m \simeq e^{-\frac{m}{2^n}}$ on $f \in_R F_n$. Therefore if $\frac{m}{2^n}$ is not close to 0, we can distinguish $f \in_R F_n$ from G' with a good probability. We will call $\mathcal{A}$ this attack. Like the attack on $\oplus_{i=1}^{2^n} G(i)$, this attack $\mathcal{A}$ requires $O(2^n)$ queries and $O(2^n)$ computations. (This attack was already described in [2].) However unlike the attack on $\oplus_{i=1}^{2^n} G(i)$, this attack $\mathcal{A}$ does not requires $m = 2^n$, but only to have $\frac{m}{2^n}$ not close to 0.

Stability of the attack

Let G'_ϕ be the function G',except on ϕ randomly and secretly chosen points x_i, and on these points G'_ϕ is perfectly random. The probability of (#) is 1 on G', is $(1 - \frac{1}{2^n})^\phi \simeq e^{\frac{\phi}{2^n}} \simeq 1 - \frac{\phi}{2^n}$ on G'_ϕ and is $\simeq e^{-\frac{m}{2^n}}$ on $f \in_R F_n$. Therefore, if ϕ is $\leq P(n)$ and if $m \simeq 2^n$, with $p(n)$ a polynomial in n, the probability of (#) is about 1 on G'_ϕ, and is about $\frac{1}{e}$ on $f \in_R F_n$, so this attack $\mathcal{A}$ is still able to distinguish G'_ϕ from $f \in_R F_n$. Therefore $\mathcal{A}$ is "stable" with our definition of "stable".

Variant with ≥ 3 Xor

With 3 Xor, instead of $G(x) = (f_1 \oplus f_2 \oplus f_3)(x)$, if $x \neq 0$, with $f_1, f_2, f_3 \in_R B_n$, and $G(0)$ random, we can study $G'(x) = f(x\|00) \oplus f(x\|01) \oplus f(x\|10)$, if $x \neq 0$, with $f \in_R B_n$ and $x \in I_{n-2}$, $G'(0)$ random. Now $G'(x)$ can have the value 0, and as with $f \oplus g \oplus h$, $f, g, h \in B_n$ with this design the best known attacks have complexity greater than $O(2^n)$. More generally, with k Xor, instead of using k random permutations of B_n, we can use only one. From a theoretical point of view the analysis, attacks and results will be similar if the number of Xor is ≥ 3

(cf Appendices D and E), but from a practical point of view these variants may be sometime a bit better since they use only one random permutation of B_n.

9 Other variants and open problems

Let assume, for example, that we want to build a pseudo-random function of F_n from two random permutations of B_n. We have

$$|B_n|^2 = \left((2^n)!\right)^2 \simeq \left((2^n)^{2^n} \cdot e^{-2^n} \sqrt{2\pi \cdot 2^n}\right)^2 \simeq 2^{2n \cdot 2^n} e^{-2 \cdot 2^n} (2\pi \cdot 2^n)$$

Here we use Stirling formula and $|F_n| = (2^n)^{2^n} = 2^{n \cdot 2^n}$. So $|B_n|^2 \geq |F_n|$ and therefore, from an information theoretic point of view, we may imagine to transform a random element of B_n^2 in a pseudo-random element of F_n with a security bound much better than $O(2^n)$. In fact, if we have a very small probability that the transformation fails, i.e. gives no element of F_n, then we may even hope to get a perfectly random element of F_n when the construction works.

Remark. A similar problem arise when we want to transform for example a perfectly random integer x of $[1, 11]$ into a perfectly random integer y of $[1, 2]$. We can decide that if $x \in \{1, 2, 3, 4, 5\}$ then $y = 1$, and if $x \in \{6, 7, 8, 9, 10\}$ then $y = 2$, and if $x = 11$, then no output y is given. Then when an output y is given, y is perfectly random in $[1, 2]$.

It may be interesting to design a similar transformation from B_n^2 to F_n, i.e. with a high probability the construction will give an output, and when it gives an output, this output will be a perfectly random element of F_n. However, we want to perform only $O(n)$ operations (or polynomial in n) to get the output (as $(f_1 \oplus g_2)(x)$ where only 2 operations are needed), not $O(2^n)$. Therefore, this problem may have no solution. However, it may exist some designs with better security results than our constructions with the same number of operations. In any case, it is an interesting and open question to evaluate the best possible designs when only $O(n)$ (or a polynomial in n) operations are possible to evaluate $G(x)$. Of course another open question is: Are our generic attacks the best possible attacks on our constructions (with k Xor and a few random points)?

10 Summary of the results

- k denotes the number of Xor: $f_1 \oplus f_2 \oplus \ldots \oplus f_k$.
- In "scenario 1" we present the number of computations required in a CPA-2 (Adaptive chosen plaintext attack) to distinguish $f_1 \oplus f_2 \oplus \ldots \oplus f_k$ (with $f_1, \ldots, f_k \in_R B_n$) from a truly random function $h \in_R F_n$ when the adversary has access to the **full** codebook. This number is proved to be at least in $O(2^n/n^{\frac{2}{3}})$ (security results of [2]) at least in $O(2^n)$ (security results of [8]), and at most in $O(2^n)$ ("unstable" attack of Section 2) when all the 2^n inputs/outputs are given), and therefore exactly in $O(2^n)$.
- "Scenario 2" is like "scenario 1" except that we have access to m input/output pairs, with $m \simeq 2^n$ but $m < 2^n$, and that we use a generator of such functions.

Table 1. Best known attacks for the Xor of k permutations

k	Scenario 1	Scenario 2 and 3	Scenario 4, $m \ll \sqrt{2^n}$	Scenario 4, $\sqrt{2^n} \ll m \ll 2^n$
2	2^n	$\leq 2^{2n}$	$Adv \geq O(\frac{m(m-1)}{2^{2n}})$	$Adv \geq O(\frac{m}{2^{\frac{3n}{2}}})$
3	2^n	$\leq 2^{4n}$	$Adv \geq O(\frac{m(m-1)}{2^{3n}})$	$Adv \geq O(\frac{m}{2^{\frac{5n}{2}}})$
4	2^n	$\leq 2^{6n}$	$Adv \geq O(\frac{m(m-1)}{2^{4n}})$	$Adv \geq O(\frac{m}{2^{\frac{7n}{2}}})$
k	2^n	$\leq 2^{(2k-2)n}$	$Adv \geq O(\frac{m(m-1)}{2^{kn}})$	$Adv \geq O(\frac{m}{2^{(k-\frac{1}{2})n}})$

- In "scenario 3" we present the number of computations required in a CPA-2 (Adaptive chosen plaintext attack) to distinguish G from a truly random function $h \in_R F_n$ where G is equal to $f_1 \oplus f_2 \oplus \ldots \oplus f_k$ (with $f_1, \ldots, f_k \in_R B_n$) on all the points except on a few points x_i where $G(x_i)$ is random. (For example it can be only on the point 0). Moreover, we use a generator of such functions G.
  "$\leq$" denotes the fact that we give here the best known attack. We see that in scenarios 2 and 3 the number of computations can be much larger than in scenario 1.Therefore the design of G can be very efficient in some applications.

- In Scenario 4 we present the best Advantage that we have found when we try to attack in CPA-2 a **single** $f_1 \oplus f_2 \oplus \ldots \oplus f_k$ with m queries, (not a generator), with $\sqrt{2^n} \ll m \ll 2^n$. (These values for Adv are always $\ll 1$, this is why in Scenarios 2 and 3 we needed more than one function to distinguish). "$\geq$" denotes the fact that we give here the best known advantage, but better Advantage may exist.

With the variant of section 8 (i.e. with only one permutation), the results obtained are the same as for $f_1 \oplus f_2 \oplus \ldots \oplus f_k$ except for $k = 2$.

Table 2. Best known attacks for the variant of section 8 (i.e. k Xor on only one permutation).

k	Scenario 1	Scenario 2 and 3	Scenario 4, $m \ll \sqrt{2^n}$	Scenario 4, $\sqrt{2^n} \ll m \ll 2^n$
2	2^n	2^n	$Adv \geq O(\frac{m}{2^n})$	$Adv \geq O(\frac{m}{2^n})$
3	2^n	$\leq 2^{4n}$	$Adv \geq O(\frac{m(m-1)}{2^{3n}})$	$Adv \geq O(\frac{m}{2^{\frac{5n}{2}}})$
4	2^n	$\leq 2^{6n}$	$Adv \geq O(\frac{m(m-1)}{2^{4n}})$	$Adv \geq O(\frac{m}{2^{\frac{7n}{2}}})$
k	2^n	$\leq 2^{(2k-2)n}$	$Adv \geq O(\frac{m(m-1)}{2^{kn}})$	$Adv \geq O(\frac{m}{2^{(k-\frac{1}{2})n}})$

11   Conclusion

In this paper, we have designed new schemes to build PRF from PRP. On these schemes we use k Xor instead of two, on all the points except a few, and on these few points, we have a truly random output. On these new schemes, we have shown that the best known generic attacks have a complexity much larger than $O(2^n)$. Therefore these schemes might be very useful when we want to generate random functions from random permutations with a small value of n and a high security (security in 2^{80} for example and $n < 80$).

References

1. William Aiello and Ramarathnam Venkatesan. Foiling Birthday Attacks in Length-Doubling Transformations - Benes: A Non-Reversible Alternative to Feistel. In Ueli M. Maurer, editor, *Advances in Cryptology – EUROCRYPT '96*, volume 1070 of *Lecture Notes in Computer Science*, pages 307–320. Springer-Verlag, 1996.
2. Mihir Bellare and Russell Impagliazzo. A Tool for Obtaining Tighter Security Analyses of Pseudorandom Function Based Constructions, with Applications to PRP to PRF Conversion. ePrint Archive 1999/024: Listing for 1999.
3. Mihir Bellare, Ted Krovetz, and Phillip Rogaway. Luby-Rackoff Backwards: Increasing Security by Making Block Ciphers Non-invertible. In Kaisa Nyberg, editor, *Advances in cryptology – EUROCRYPT 1998*, volume 1403 of *Lecture Notes in Computer Science*, pages 266–280. Springer-Verlag, 1998.
4. Chris Hall, David Wagner, John Kelsey, and Bruce Schneier. Building PRFs from PRPs. In Hugo Krawczyk, editor, *Advances in Cryptology – CRYPTO 1998*, volume 1462 of *Lecture Notes in Computer Science*, pages 370–389. Springer-Verlag, 1998.
5. Stefan Lucks. The Sum of PRPs Is a Secure PRF. In Bart Preneel, editor, *Advances in Cryptology – EUROCRYPT 2000*, volume 1807 of *Lecture Notes in Computer Science*, pages 470–487. Springer-Verlag, 2000.
6. Avradip Mandal, Jacques Patarin, and Valérie Nachef. Indifferentiability beyond the Birthday Bound for the Xor of Two Public Random Permutations. In Guang Gong and Kishan Chand Gupta, editors, *Progress in Cryptology – INDOCRYPT 2010*, volume 6948 of *Lecture Notes in Computer Science*, pages 69–81. Springer-Verlag, 2010.
7. Ueli Maurer and Krzysztof Pietrzak. The Security of Many-Round Luby-Rackoff Pseudo-Random Permutations. In Eli Biham, editor, *Advances in Cryptology – EUROCRYPT 2003*, volume 2656 of *Lecture Notes in Computer Science*, pages 544–561. Springer-Verlag, 2003.
8. J. Patarin. A Proof of Security in $O(2^n)$ for the Xor of Two Random Permutations — Proof with the H_σ technique. . In R. Safavi, editor, *ICITS '2008*, volume 5155 of *Lecture Notes in Computer Science*, pages 232–248. Springer-Verlag, 2008.
9. Jacques Patarin. Security in $O(2^n)$ for the Xor of Two Random Permutations — Proof with the standard H technique. This paper is available from the author.

Appendices

A Mean value and standard deviation of collisions on random functions

Aim. Let f be a random function from I_n to I_n. We assume that we know f on m distinct points x_i: $\forall i,\ 1 \leq i \leq m,\ f(x_i) = y_i$. Let N be the number of collisions on these values y_i. We want to evaluate $E(N)$ (the mean value of N when $f \in_R F_n$) and $\sigma(N)$ (the standard deviation of N when $f \in_R F_n$).

Computation of $E(N)$. Let $\delta_{ij} = 1 \Leftrightarrow f(x_i) = f(x_j)$ and $\delta_{ij} = 0 \Leftrightarrow \delta_{ij} \neq 1$. We have $N = \sum_{i<j} \delta_{ij}$. Therefore, $E(N) = \sum_{i<j} E(\delta_{ij})$. Moreover

$$E(\delta_{ij}) = Pr_{\substack{i \neq j \\ f \in_R B_n}} \left(f(x_i) = f(x_j) \right) = \frac{1}{2^n}$$

Therefore $E(N) = \frac{m(m-1)}{2 \cdot 2^n}$.

Computation of $\sigma(N)$.

$$V(N) = V\left(\sum \delta_{ij}\right) = \sum_{i<j} V(\delta_{ij}) + \sum_{\substack{i<j,\ k<l, \\ (i,j) \neq (k,l)}} Cov(\delta_{ij}, \delta_{kl})$$

where $Cov(\delta_{ij}, \delta_{kl})$ denotes the covariance of $(\delta_{ij}, \delta_{kl})$:

$$Cov(\delta_{ij}, \delta_{kl}) = E(\delta_{ij} \cdot \delta_{kl}) - E(\delta_{ij})E(\delta_{kl})$$

We have:

$$V(\delta_{ij}) = E(\delta_{ij}^2) - E(\delta_{ij})^2 = \frac{1}{2^n} - \frac{1}{2^{2n}}$$

We now have to evaluate $E(\delta_{ij} \cdot \delta_{kl})$.

Case 1: i, j, k, l are pairwise distinct. Then

$$E(\delta_{ij} \cdot \delta_{kl}) = Pr_{f \in_R B_n}\left(f(x_i) = f(x_j) \text{ and } f(x_k) = f(x_l) \right) = \frac{1}{2^{2n}}$$

Case 2: In i, j, k, l, we have exactly 3 distinct values. For example $i = k$. Then

$$E(\delta_{ij} \cdot \delta_{kl}) = Pr_{f \in_R B_n}\left(f(x_i) = f(x_j) = f(x_l) \right) = \frac{1}{2^{2n}}$$

Therefore all the covariance are 0 and we have:

$$V(N) = \frac{m(m-1)}{2}\left(\frac{1}{2^n} - \frac{1}{2^{2n}}\right) \quad \text{and} \quad \sigma(N) = \sqrt{V(N)} = O\left(\frac{m}{\sqrt{2^n}}\right)$$

B Mean value and standard deviation of collisions on $f \oplus g,\ f, g \in_R B_n$

Aim. Let $G = f \oplus g$, with $f, g \in_R B_n$. We assume that we know G on m distinct points x_i: $\forall i,\ 1 \leq i \leq m,\ G(x_i) = y_i$. Let N be the number of collisions on these m values y_i. We want to evaluate $E(N)$ (the mean value of N when $f, g \in_R B_n$) and $\sigma(N)$ (the standard deviation of N when $f, g \in_R B_n$).

Computation of $E(N)$. Let $\delta_{ij} = 1 \Leftrightarrow G(x_i) = G(x_j)$ and $\delta_{ij} = 0 \Leftrightarrow \delta_{ij} \neq 1$. We have $N = \sum_{i<j} \delta_{ij}$. Therefore, $E(N) = \sum_{i<j} E(\delta_{ij})$. Moreover

$$E(\delta_{ij}) = Pr_{\substack{i \neq j, \\ f,g \in_R B_n}} \left(g(x_i) \oplus g(x_j) = f(x_i) \oplus f(x_j) \right)$$

When f is fixed, $f \in B_n$, $f(x_i) \oplus f(x_j)$ is a value different from 0. Therefore the probability when $g \in_R B_n$ that $g(x_i) \oplus g(x_j) = f(x_i) \oplus f(x_j)$ is exactly $\frac{1}{2^n-1}$. So

$$E(\delta_{ij}) = \frac{1}{2^n - 1} \quad \text{and} \quad E(N) = \frac{m(m-1)}{2} \cdot \frac{1}{2^n - 1}$$

Computation of $\sigma(N)$.

$$V(N) = V\Big(\sum_{i<j} \delta_{ij}\Big) = \sum_{i<j} V(\delta_{ij}) + \sum_{\substack{i<j,\ k<l \\ (i,j) \neq (k,l)}} Cov(\delta_{ij}, \delta_{kl}) \quad (*)$$

where $Cov(\delta_{ij}, \delta_{kl})$ denotes the covariance of $(\delta_{ij}, \delta_{kl})$:

$$Cov(\delta_{ij}, \delta_{kl}) = E(\delta_{ij} \cdot \delta_{kl}) - E(\delta_{ij})E(\delta_{kl})$$

We have:

$$V(\delta_{ij}) = E(\delta_{ij}^2) - E(\delta_{ij})^2 = \frac{1}{2^n - 1} - \frac{1}{(2^n - 1)^2}$$

We now have to evaluate $E(\delta_{ij} \cdot \delta_{kl})$

Case 1: i, j, k, l are pairwise distinct. Then

$$E(\delta_{ij} \cdot \delta_{kl}) = Pr_{f,g \in_R B_n} \left(\begin{array}{l} g(x_i) \oplus g(x_j) = f(x_i) \oplus f(x_j) \\ g(x_k) \oplus g(x_l) = f(x_k) \oplus f(x_l) \end{array} \right)$$

When $f(x_i)$, $f(x_j)$, $f(x_k)$, $f(x_l)$, $g(x_j)$, $g(x_l)$ are fixed, $g(x_i)$ and $g(x_k)$ are fixed with

$$g(x_i) = g(x_j) \oplus f(x_i) \oplus f(x_j) \quad \text{and} \quad g(x_k) = g(x_l) \oplus f(x_k) \oplus f(x_l)$$

(and these conditions may be compatible or not with g being a permutation). If we did not have these two equalities, for $g(x_i)$ we would have $(2^n - 2)$ possibilities ($g(x_i) \notin \{g(x_j), g(x_l)\}$), and for $g(x_k)$ we would have $(2^n - 3)$ possibilities ($g(x_k) \notin \{g(x_i), g(x_j), g(x_l)\}$). So,

$$E(\delta_{ij} \cdot \delta_{kl}) \leq \frac{1}{(2^n - 2)(2^n - 3)}$$

Therefore

$$E(\delta_{ij} \cdot \delta_{kl}) - E(\delta_{ij})E(\delta_{kl}) \leq \frac{1}{(2^n - 2)(2^n - 3)} - \frac{1}{(2^n - 1)^2}$$

$$\leq \frac{3 \cdot 2^n}{(2^n - 1)^2(2^n - 2)(2^n - 3)} \leq O(\frac{1}{2^{3n}})$$

Case 2: in i, j, k, l, we have exactly 3 distinct values. For example $i = k$. Then

$$E(\delta_{ij} \cdot \delta_{kl}) = Pr_{f,g \in B_n}(f(x_i) \oplus g(x_i) = f(x_j) \oplus g(x_j) = f(x_l) \oplus g(x_l))$$

When $f(x_i)$, $f(x_j)$, $f(x_l)$, $g(x_i)$ are fixed, $g(x_j)$ and $g(x_l)$ are fixed with

$$\begin{cases} g(x_j) = f(x_i) \oplus g(x_i) \oplus f(x_j) \\ g(x_l) = f(x_i) \oplus g(x_i) \oplus f(x_l) \end{cases}$$

(and these conditions may be compatible or not with g being a permutation). If we did not have these two equalities, for $g(x_j)$ we would have $(2^n - 1)$ possibilities ($g(x_j) \neq g(x_i)$) and for $g(x_l)$, we would have $(2^n - 2)$ possibilities ($g(x_l) \notin \{g(x_i), g(x_j)\}$). So

$$E(\delta_{ij} \cdot \delta_{kl}) \leq \frac{1}{(2^n - 1)(2^n - 2)}$$

Therefore

$$E(\delta_{ij} \cdot \delta_{kl}) - E(\delta_{ij})E(\delta_{kl}) \leq \frac{1}{(2^n - 1)(2^n - 2)} - \frac{1}{(2^n - 1)^2}$$

$$\leq \frac{1}{(2^n - 1)^2(2^n - 2)} \leq O(\frac{1}{2^{3n}})$$

So from $(*)$ we get

$$V(N) \leq \frac{m(m-1)}{2}\left(\frac{1}{2^n - 1} - \frac{1}{(2^n - 1)^2}\right) + O(\frac{m^4}{2^{3n}})$$

So

$$V(N) \leq O(\frac{m^2}{2^n}) + O(\frac{m^4}{2^{3n}})$$

Since $m \leq 2^n$, $V(N) \leq O(\frac{m^2}{2^n})$ and therefore $\sigma(N) \leq O(\frac{m}{\sqrt{2^n}})$.

C Mean value and standard deviation of collisions on $f_1 \oplus f_2 \oplus \ldots \oplus f_k$

Theorem 1 *Let* $G = f_1 \oplus f_2 \oplus \ldots \oplus f_k$, $f, g \in_R B_n$, *with* $f_1, f_2, \ldots, f_k \in_R B_n$. *Let assume that we know* G *on* m *distinct points* x_i: $\forall i$, $1 \leq i \leq m$, $G(x_i) = y_i$. *Let* N_k *be the number of collisions on these* m *points:* N_k = *the number of* (i, j), $1 \leq i < j \leq m$ *such that* $y_i = y_j$. *Then*

$$E(N_k) = \frac{m(m-1)}{2} \cdot \frac{1}{2^n}\left[1 + \frac{(-1)^k}{(2^n - 1)^{k-1}}\right]$$

where $E(N_k)$ *denotes the mean value of* N_k *when* $f_1, f_2, \ldots, f_k$ *are randomly chosen in* B_n.

To prove this theorem we will first need a lemma.

Lemma 1. *If $x_i \neq x_j$, we have*

$$\text{if } \varphi \neq 0, \quad Pr_{f \in B_n}\left(f(x_i) \oplus f(x_j) = \varphi\right) = \frac{1}{2^n - 1}$$

$$\text{and if } \varphi = 0, \quad Pr_{f \in B_n}\left(f(x_i) \oplus f(x_j) = \varphi\right) = 0$$

Proof of Lemma 1

If $\varphi = 0$, $f(x_i) \neq f(x_j)$ since f is a permutation. If $\varphi \neq 0$, when $f(x_i)$ is fixed, $f(x_j)$ is fixed to the value of $\varphi \oplus f(x_i)$, so instead of having $2^n - 1$ possible values for $f(x_j)$ we have one when $f(x_i)$ is fixed.

Proof of Theorem 1

Let $\delta_{ij} = 1 \Leftrightarrow G(x_i) = G(x_j)$ and $\delta_{ij} = 0 \Leftrightarrow \delta_{ij} \neq 1$. We have $N_k = \sum_{i<j} \delta_{ij}^k$, so $E(N_k) = \sum_{i<j} E(\delta_{ij}^k)$. We will compute $E(\delta_{ij}^k)$ by induction on k.

$$E(\delta_{ij}^k) = Pr_{f_1,\ldots,f_k \in_R B_n}\left[f_1(x_i) \oplus \ldots \oplus f_k(x_i) = f_1(x_j) \oplus \ldots \oplus f_k(x_j)\right]$$

So from Lemma 1 above,

$$E(\delta_{ij}^k) = \frac{1}{2^n - 1} Pr_{f_1,\ldots,f_{k-1} \in_R B_n}\left[f_1(x_i) \oplus \ldots \oplus f_{k-1}(x_i) \neq f_1(x_j) \oplus \ldots \oplus f_{k-1}(x_j)\right]$$

$$E(\delta_{ij}^k) = \frac{1}{2^n - 1}\left[1 - E(\delta_{ij}^{k-1})\right] \quad (*)$$

If $k = 1$ we have $E(\delta_{ij}^1) = Pr_{f_1 \in B_n}(f_1(x_i) = f_1(x_j)) = 0 \quad (**)$ (since f_1 is a permutation and $x_i \neq x_j$). Now from $(*)$ and $(**)$ we get immediately by induction on k that

$$E(\delta_{ij}^k) = \frac{1}{2^n}\left[1 + \frac{(-1)^k}{(2^n - 1)^{k-1}}\right]$$

and therefore,

$$E(N_k) = \frac{m(m-1)}{2} E(\delta_{ij}^k) = \frac{m(m-1)}{2} \cdot \frac{1}{2^n}\left[1 + \frac{(-1)^k}{(2^n - 1)^{k-1}}\right]$$

as claimed. Moreover the standard deviation can be computed exactly as in Appendix B, or alternatively by using the fact that $G = f_1 \oplus f_2 \oplus \psi$ where ψ is a function independant of $f_1 \oplus f_2$. We get the same result: $\sigma(N_k) \leq O(\frac{m}{\sqrt{2^n}})$.

Remark. This result is not surprising: by Xoring k permutations, $k \geq 3$ instead of 2, we expect to obtain a better or at least as good pseudorandom permutation. Since we have seen that $\sigma(N)$ for $k = 2$ and $\sigma(N)$ for a random function are less than or equal to $O(\frac{m}{\sqrt{2^n}})$, it is natural that for $k \geq 3$ we also have the same result $\sigma(N) \leq O(\frac{m}{\sqrt{2^n}})$.

D   Mean value of collisions on $f(x\|\alpha) \oplus f(x\|\beta)$, $f \in_R B_n$

Let $G'(x) = f(x\|\alpha) \oplus f(x\|\beta)$, $f \in_R B_n$, with $\alpha \neq \beta$. We assume that we know G' on m distinct points x_i: $\forall i$, $1 \leq i \leq m$, $G'(x_i) = y_i$. Let N be the number of collisions on these m values y_i. We want to evaluate $E(N)$, the mean value of N when $f \in_R B_n$. Let $\delta_{ij} = 1 \Leftrightarrow G'(x_i) = G'(x_j)$ and $\delta_{ij} = 0 \Leftrightarrow \delta_{ij} \neq 1$. We have $N = \sum_{i<j} \delta_{ij}$. Therefore $E(N) = \sum_{i<j} E(\delta_{ij})$. Moreover $E(\delta_{ij}) = Pr_{f \in_R B_n}(f(x_i\|\alpha) \oplus f(x_i\|\beta) \oplus f(x_j\|\alpha) = f(x_j\|\beta))$. So $E(\delta_{ij}) = Pr_{f \in_R B_n}(f(a) \oplus f(b) \oplus f(c) = f(d))$ where a, b, c, d are pairwise distinct. When $f(a)$, $f(b)$ and $f(b)$ are fixed, then $f(d)$ can have any value $\notin \{f(a), f(b), f(c)\}$ with probability exactly $\frac{1}{2^n-3}$ (and $f(d) \in \{f(a), f(b), f(c)\}$ with probability 0). Moreover $f(a) \oplus f(b) \oplus f(c) \in \{f(a), f(b), f(c)\}$ is not possible since f is a permutation. Therefore $E(\delta_{ij}) = \frac{1}{2^n-3}$ and $E(N) = \frac{m(m-1)}{2.(2^n-3)} \simeq \frac{m(m-1)}{2.2^n}(1 + \frac{3}{2^n})$.

E   Mean value of collisions on $f(x\|\alpha_1) \oplus \ldots \oplus f(x\|\alpha_k)$, $f \in_R B_n$

Let $G'_k(x) = f(x\|\alpha_1) \oplus \ldots \oplus f(x\|\alpha_k)$, $f \in_R B_n$, with $\alpha_1, \alpha_2, \ldots \alpha_k$ pairwise distinct. We assume that we know G'_k on m distinct points x_i: $\forall i$, $1 \leq i \leq m$, $G'_k(x_i) = y_i$. Let N_k be the number of collisions on these m values y_i. We want to evaluate $E(N_k)$, the mean value of N when $f \in_R B_n$. Let $\delta_{ij} = 1 \Leftrightarrow G'_k(x_i) = G'_k(x_j)$ and $\delta_{ij} = 0 \Leftrightarrow \delta_{ij} \neq 1$. We have $N_k = \sum_{i<j} \delta_{ij}$. Therefore $E(N_k) = \sum_{i<j} E(\delta_{ij})$. Let $p_k = E(\delta_{ij}) = Pr_{f \in_R B_n}(f(x_i\|\alpha_1) \oplus \ldots \oplus f(x_i\|\alpha_k) = f(x_j\|\alpha_1) \oplus \ldots \oplus f(x_j\|\alpha_k)) = Pr_{f \in_R B_n}(f(a_1) \oplus f(a_2) \oplus \ldots \oplus f(a_{2k-1}) = f(a_{2k}))$ where $a_1, a_2, \ldots, a_{2k}$ are pairwise distinct. When $f(a_1), \ldots f(a_{2k-1})$ are fixed, then $f(a_{2k})$ can have any value $\notin \{f(a_1), \ldots, f(a_{2k-1})\}$ with probability exactly $\frac{1}{2^n-(2k-1)}$ (and $f(a_{2k}) \in \{f(a_1), \ldots, f(a_{2k-1})\}$ with probability 0). Therefore we have: $p_k = (1 - (2k-1)p_{k-1}).\frac{1}{2^n-(2k-1)}$ $(*)$ (since $\forall i$, $1 \leq i \leq 2k-1$ we have the probability exactly $1 - p_{k-1}$ that $f(a_1) \oplus f(a_2) \oplus \ldots f(a_{2k-1}) = f(a_i)$). For example, from $p_1 = 0$ (since f is a bijection), we get from $(*)$: $p_2 = \frac{1}{2^n-3}$ (as already found in Appendix D), and then $p_3 = (1 - 5p_2).\frac{1}{2^n-5} = \frac{1}{2^n}(1 - \frac{15}{2^{2n}-8.2^n+15})$. More generally, from $(*)$ and $p_2 = \frac{1}{2^n-3}$, we get easily by induction that:
$p_k = \frac{1}{2^n}\left[1 + \frac{(-1)^k.3.5.7\ldots(2k-1)}{2^{(k-1)n}(1-\frac{3}{2^n})(1-\frac{5}{2^n})\ldots(1-\frac{2k-1}{2^n})}\right]$. Therefore $E(N_k) = \frac{m(m-1)}{2}p_k = \frac{m(m-1)}{2.2^n}\left[1 + O(\frac{1}{2^{(k-1)n}})\right]$, with $O(\frac{1}{2^{(k-1)n}}) = \frac{(-1)^k.3.5.7\ldots(2k-1)}{2^{(k-1)n}(1-\frac{3}{2^n})(1-\frac{5}{2^n})\ldots(1-\frac{2k-1}{2^n})}$.