J. Math. Cryptol. **4** (2010), 239–270
DOI 10.1515/JMC.2010.010

Multicollision attacks and generalized iterated hash functions

Juha Kortelainen, Kimmo Halunen and Tuomas Kortelainen

Communicated by Douglas R. Stinson

Abstract. We apply combinatorics on words to develop an approach to multicollisions in generalized iterated hash functions. Our work is based on the discoveries of A. Joux and on generalizations provided by M. Nandi and D. Stinson as well as J. Hoch and A. Shamir. We wish to unify the existing diverse notation in the field, bring basic facts together, reprove some previously published results and produce some new ones. A multicollision attack method informally described by Hoch and Shamir is laid on a sound statistical basis and studied in detail.

Keywords. Hash functions, combinatorics on words, multicollision.

2010 Mathematics Subject Classification. 94A60, 68R15.

1 Introduction

Iterated hash functions have been the most successful method for constructing fast and secure hash functions. The underlying principle proposed by Merkle and Damgård [4, 15] is quite simple and easy to implement. However, most of the modern hash functions built on this foundation were proved insecure in [7, 12, 17, 19, 20]. Many of these flaws come from the weaknesses in the underlying compression functions. In recent years, more rigorous theoretical study has also found some weaknesses in the iterative structure itself [3].

One of the most notable results on the iterative structure was Joux's method concerning multicollisions in iterated hash functions [10], which has been used to disprove some of the assumptions on hash function security. Furthermore, these achievements concerning multicollisions were generalized by Nandi and Stinson [16] and later by Hoch and Shamir [9]. These results show that Joux's method can be applied to a more general class of iterated hash functions.

The theoretical results concerning hash functions and especially multicollisions in iterated hash functions were created with a multitude of different approaches, notation and a varying level of mathematical rigor. This has made it somewhat

Some preliminary results of this research have been published in AISC 2010 proceedings [8].

difficult to examine the differences and similarities of the results achieved. It has also led to the situation where we do not have a unified theoretical and notational framework for the study of iterated hash functions in general and multicollisions in particular.

In this paper, we show a way to formulate these problems in a well-established mathematical system and use this structure to prove the central results related to multicollisions. The notation and basic theory of combinatorics on words, algebra and partial orders is extensively applied. In [9] a method of constructing multicollisions in Iterated Concatenated and Expanded (ICE) hash functions is introduced. The description of the method is informal and difficult, if not impossible, to understand in detail. We wish to give a rigorous mathematical treatment to this method and point out certain deficiencies in the original version of it. A fairly detailed complexity analysis of the respective attack construction is provided and the practical applicability of the attack and the ICE hash functions is discussed.

The paper is organized in the following way. The second section introduces the basic definitions of combinatorics on words and partial orders. The third section shows how iterated hash functions and multicollisions can be depicted in this theoretical setting; the earlier work conducted in the field is reviewed and the structure of the attack schema on generalized iterated hash functions is described informally. In the fourth section, we prove the combinatorial results needed for the construction of a feasible multicollision attack on so-called bounded generalized iterated hash functions. The fifth section contains the precise exposition of the multicollision attack and some analysis of its complexity. In the final section, we discuss our results, draw some conclusions from our research and give possible future research proposals.

2 Basics on words, languages and partial orders

We encourage the reader to return to the basic concepts only as the need arises.

2.1 Words and languages

Let $\mathbb{N} = \{0, 1, 2, \ldots\}$ be the set of all natural numbers and $\mathbb{N}_+ = \mathbb{N}\backslash\{0\}$. For each $l \in \mathbb{N}_+$, we define $\mathbb{N}_l$ to be the set of l first positive integers: $\mathbb{N}_l = \{1, 2, \ldots, l\}$. For each finite set S, let $|S|$ be the cardinality of S, i.e., the number of elements in S.

An *alphabet* is any finite nonempty set of abstract symbols called *letters*. Let A be an alphabet. A *word* (over A) is any finite sequence of symbols in A. Thus, assuming that w is a word over A, we can write $w = a_1 a_2 \cdots a_n$, where $n \in \mathbb{N}$ and $a_i \in A$ for $i = 1, 2, \ldots, n$. Here n is the *length* $|w|$ of w. Notice that n may be equal to zero; then w is the *empty word*, denoted by ϵ, that contains

no letters. By $|w|_a$ we mean the number of occurrences of the letter a in w. Denote $\mathrm{alph}(w) = \{a \in A \,|\, |w|_a > 0\}$. Obviously $\mathrm{alph}(\epsilon) = \emptyset$; for nonempty w, call $\mathrm{alph}(w)$ the *alphabet of* w. Let A^* (resp. A^+) be the set of all words (resp. nonempty words) over A. By A^n, $n \in \mathbb{N}_+$, we mean the set of all words of length n over A. The *catenation* of two words u and v in A^* is the word uv obtained by writing v after u. Clearly, catenation defines a binary operation $\cdot$ in A^*: $u \cdot v = uv$ for all $u, v \in A^*$. In algebraic terms $(A^*, \cdot)$ is a *free monoid* and $(A^+, \cdot)$ is a *free semigroup*. The word u is a *subword* of w if there exist $m \in \mathbb{N}$ and $x_0, u_1, x_1, \ldots, u_m, x_m \in A^*$ such that $w = x_0 u_1 x_1 \cdots u_m x_m$ where $u = u_1 u_2 \cdots u_m$. A subword u of w is a *factor* of w if $w = x_0 u x_1$ for some $x_0, x_1 \in A^*$. If a word y can be written in the form $y = y_1 y_2 \cdots y_p$ for some $p \in \mathbb{N}$ and words $y_1, y_2, \ldots, y_p$, we say that $y_1 y_2 \cdots y_p$ is a *factorization* of y (into $y_1, y_2, \ldots, y_p$).

Let A and B be alphabets. A mapping $h : A^* \to B^*$ is a (*monoid*) *morphism* if $h(uv) = h(u)h(v)$ for each $u, v \in A^*$. Note that a morphism always maps the empty word ϵ to ϵ. Moreover, the morphism h is completely determined by the images $h(a)$ of all letters $a \in A$. If $B \subseteq A$, then the *projection morphism* from A^* into B^*, denoted by π_B^A (or π_B, when A is understood), is defined by $\pi_B^A(b) = b$ for each $b \in B$ and $\pi_B^A(a) = \epsilon$ for each $a \in A \setminus B$.

A *permutation* of an alphabet A is any word $w \in A^+$ such that $|w|_a = 1$ for each $a \in A$.

A *language* (over the alphabet A) is any set of words L (such that $L \subseteq A^*$). Let L_1 and L_2 be languages. The *catenation* of L_1 and L_2 is the language $L_1 L_2 = \{uv \,|\, u \in L_1, v \in L_2\}$. Define the *powers* of the L_1 recursively as follows: $L_1^1 = L_1$, and $L_1^{i+1} = L_1^i L_1$ for $i \in \mathbb{N}_+$. The *positive closure* of L_1 is the language $L_1^+ = \bigcup_{i=1}^{\infty} L_1^i$. For any word w, we write w^+ instead of $\{w\}^+$.

2.2 Partial orders

A binary relation R on the nonempty set X is a *partial order* (*in* X) if it is irreflexive ($\forall x \in X : (x, x) \notin R$), antisymmetric ($\forall x, y \in X : (x, y) \in R \Rightarrow (y, x) \notin R$) and transitive ($\forall x, y, z \in X : (x, y) \in R \wedge (y, z) \in R \Rightarrow (x, z) \in R$).

Let $\prec$ be a partial order in X. Call $(X, \prec)$ a *partially ordered set*. The elements $x, y \in X$, $x \neq y$, are *incomparable* (in $(X, \prec)$) if neither $x \prec y$ nor $y \prec x$ holds. The nonempty finite sequence $x_1, x_2, \ldots, x_n$ of elements of X is a *chain* of $(X, \prec)$ if $x_i \prec x_{i+1}$ for all $i \in \{1, 2, \ldots, n-1\}$. Above $n \in \mathbb{N}_+$ is the *length* of the chain $x_1 \prec x_2 \prec \cdots \prec x_n$. For each chain c of $(X, \prec)$, let $|c|$ be the length of c. An (indexed) set of chains $\{c_i\}_{i \in I}$ is a *chain decomposition* of $(X, \prec)$, if $\{C_i\}_{i \in I}$ is a partition of X, where $C_i = \{x \in X \,|\, x \text{ occurs in the chain } c_i\}$. Obviously, a chain decomposition exists for all partially ordered sets.

Now consider a finite partially ordered set $(X, \prec)$, i.e., a partially ordered set such that X is finite. The *maximum number of incomparable elements* of $(X, \prec)$ is the cardinality of the largest set $Y \subseteq X$ such that the elements of Y are pairwise incomparable. The *minimum chain decomposition size* of $(X, \prec)$ is the smallest number $m \in \mathbb{N}_+$ such that there exist chains $c_1, c_2, \ldots, c_m$ of $(X, \prec)$ for which $\{c_i\}_{i=1}^m$ is a chain decomposition of $(X, \prec)$. Finally, let the *maximum chain length* of $(X, \prec)$ be the greatest number $m \in \mathbb{N}_+$ such that there exists a chain of length m in $(X, \prec)$.

An important connection between the first two concepts defined above is stated in a famous theorem of Dilworth [6].

Theorem 2.1 (Dilworth's Theorem). *Let $(X, \prec)$ be a finite, partially ordered set. Then the maximum number of incomparable elements of $(X, \prec)$ is equal to the minimum chain decomposition size of $(X, \prec)$.*

Let us now investigate partial orders induced by words. Let α be a nonempty word. Define the binary relation $\prec_\alpha$ on $\mathrm{alph}(\alpha)$ as follows. For each $a, b \in \mathrm{alph}(\alpha)$, let $a \prec_\alpha b$ hold if and only if $a \neq b$ and all occurrences of a in α lie before any occurrence of b in α. Certainly if $a \prec_\alpha b$, then there exist words α_1 and α_2 such that $\alpha = \alpha_1\alpha_2$ and $|\alpha_1|_b = |\alpha_2|_a = 0$. Obviously, $\prec_\alpha$ is irreflexive, antisymmetric and transitive, so $(\mathrm{alph}(\alpha), \prec_\alpha)$ is a partially ordered set. Call the elements of a nonempty set $A \subseteq \mathrm{alph}(\alpha)$ *independent* (with respect to $\prec_\alpha$) if they form a chain in $(\mathrm{alph}(\alpha), \prec_\alpha)$. Now suppose that A consists of $k \in \mathbb{N}_+$ independent elements. There then exist elements $a_1, a_2, \ldots, a_k$ of $\mathrm{alph}(\alpha)$ such that $a_1 \prec_\alpha a_2 \prec_\alpha \cdots \prec_\alpha a_k$ and $A = \{a_1, a_2, \ldots, a_k\}$. Certainly $\pi_A(\alpha) \in a_1^+ a_2^+ \cdots a_k^+$.

The partial order $\prec_\alpha$ plays a central role in the construction of multicollisions as well as in their combinatorial analysis. The role of partial orders applied to combinatorics on words is extensively studied in [5].

3 Hash functions and collisions

In this section, we give the basic definitions of hash functions and multicollisions using a fresh and rigorous notation. The principles of (iterative) hash functions were, however, presented already in [4], and advanced ideas on multicollisions appear in [9, 10, 16].

3.1 Fundamental concepts

By a *block representation* of a message, we mean the division and padding of the message into blocks of equal size. We may certainly assume, without loss of

generality, that all our messages are written in the binary alphabet $\{0,1\}$ and given in a block representation form.

Definition 3.1. A *hash function of length n* (with $n \in \mathbb{N}_+$) is a mapping $f : \{0,1\}^* \to \{0,1\}^n$.

An ideal hash function $f : \{0,1\}^* \to \{0,1\}^n$ is a *variable input length random oracle* (VIL-RO for short): for each $x \in \{0,1\}^*$, the value $f(x) \in \{0,1\}^n$ is chosen uniformly at random.

Let f be a hash function. *A preimage* of a given hash value y is any $x \in \{0,1\}^*$ such that $f(x) = y$. A *second preimage* of $y = f(x)$ is any $x' \in \{0,1\}^*$ such that $f(x') = y$ and $x \neq x'$.

Let $k \in \mathbb{N}_+$. A *k-collision* in the hash function $f : \{0,1\}^* \to \{0,1\}^n$ is a set $A \subseteq \{0,1\}^*$ such that $|A| = k$ and $f(x) = f(y)$ for all $x, y \in A$. A 2-collision (in f) is also called a collision in f.

A *k-collision attack* on a hash function f can be loosely characterized as a probabilistic procedure (based on the birthday paradox) that finds a k-collision in f with some nonnegligible probability. The *complexity* of the attack can be measured, for instance, with respect to the expected number of messages the hash values of which have to be determined in order to carry out the attack successfully.

According to the (generalized) *birthday paradox*, a k-collision can be found (with probability approx. 0.5) by hashing $(k!)^{\frac{1}{k}} 2^{\frac{n(k-1)}{k}}$ different messages [18]. In the case $k = 2$ this gives $\sqrt{2} \cdot 2^{\frac{n}{2}}$ hash function computations. Intuitively most of us would expect the number to be around 2^{n-1}. For preimages and second preimages the complexity of an attack on an ideal hash function is in $O(2^n)$.

In the following, we shall derive the concept of a (generalized) iterated hash function. Remember that all messages are assumed to be in a block representation form.

Definition 3.2. A *compression function* (of block size m and length n) is a mapping $f : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}^n$ where $m, n \in \mathbb{N}_+, m > n$.

Again, an ideal compression function $f : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}^n$ is a *fixed input length random oracle* (FIL-RO for short): for each $h \in \{0,1\}^n$ and $y \in \{0,1\}^m$, the value $f(h, y) \in \{0,1\}^n$ is chosen uniformly at random.

Let $m, n \in \mathbb{N}_+, m > n$ and $f : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}^n$ be a given compression function.

Define the function $f^+ : \{0,1\}^n \times (\{0,1\}^m)^+ \to \{0,1\}^n$ inductively as follows. Let $h \in \{0,1\}^n$, $y_1 \in \{0,1\}^m$, and $y_2 \in (\{0,1\}^m)^+$. Then $f^+(h, y_1) = f(h, y_1)$

and $f^+(h, y_1y_2) = f^+(f(h, y_1), y_2)$. Surely for all $y, y' \in (\{0,1\}^m)^+$, the equality $f^+(h, y\, y') = f^+(f^+(h, y), y')$ holds.

Let u be a word in $(\{0,1\}^m)^+$ such that $u = u_1u_2\cdots u_l$ where $l \in \mathbb{N}_+$ and $u_i \in \{0,1\}^m$ for $i = 1, 2, \ldots, l$. Define the morphism $\bar{u} : \mathbb{N}_l^* \to \{0,1\}^*$ by $\bar{u}(i) = u_i$ for each $i \in \mathbb{N}_l$. Let $\alpha \in \mathbb{N}_l^+$ be given. Then $\alpha = i_1i_2\cdots i_s$, where $s \in \mathbb{N}_+$ and $i_j \in \mathbb{N}_l$ for $j = 1, 2, \ldots, s$. By definition, $\bar{u}(\alpha) = u_{i_1}u_{i_2}\cdots u_{i_s}$. Obviously $\bar{u}(\alpha)$ is the word where blocks taken from $u_1, u_2, \ldots, u_l$ are written in the order and multiple determined by α.

Example 3.3. Let $u = u_1u_2\cdots u_5$ where $u_i \in \{0,1\}^m$ for $i = 1, 2, 3, 4, 5$. Then $\bar{u}(2\cdot 5\cdot 1\cdot 3\cdot 5\cdot 5\cdot 1) = u_2u_5u_1u_3u_5u_5u_1$.

Now define the *iterated compression function* $f_\alpha : \{0,1\}^n \times \{0,1\}^{ml} \to \{0,1\}^n$ (based on α and f) by $f_\alpha(h, u) = f^+(h, \bar{u}(\alpha))$ for each $h \in \{0,1\}^n$ and $u \in \{0,1\}^{ml}$.

It is clear from the definitions that if $\alpha = \alpha_1\alpha_2$, where $\alpha_1, \alpha_2 \in \mathbb{N}_l^+$, then

$$\begin{aligned} f_\alpha(h, u) &= f^+(h, \bar{u}(\alpha)) = f^+(h, \bar{u}(\alpha_1)\bar{u}(\alpha_2)) \\ &= f^+(f^+(h, \bar{u}(\alpha_1)), \bar{u}(\alpha_2)) = f_{\alpha_2}(f_{\alpha_1}(h, u), u) \end{aligned}$$

for each $h \in \{0,1\}^n$ and $u \in \{0,1\}^{ml}$. Given $k \in \mathbb{N}_+$ and $h_0 \in \{0,1\}^m$, a *k-collision (with initial value h_0) in the iterated compression function f_α* is a set $A \subseteq \{0,1\}^{ml}$ such that $|A| = k$ and $f_\alpha(h_0, u) = f_\alpha(h_0, v)$ for all $u, v \in A$. We say that the k-collision A in f_α is *nontrivial* if, for each $u = u_1u_2\cdots u_l$ and $v = v_1v_2\cdots v_l$ in A such that $u_i, v_i \in \{0,1\}^m$ for $i = 1, 2, \ldots, l$, the equality $u_j = v_j$ holds for each $j \in \mathbb{N}_l \setminus \mathrm{alph}(\alpha)$.

Example 3.4. Let $l = 5$ and $\alpha = 3\cdot 1\cdot 5\cdot 4\cdot 4\cdot 3\cdot 1$ a word over the alphabet $\mathbb{N}_5$. Assume furthermore that $h_0 \in \{0,1\}^n$ and $u_1, u_2, u_3, u_4, u_5, u_5' \in \{0,1\}^m$, $u_5 \neq u_5'$ are message blocks such that $f^+(h_0, u_3u_1u_5) = f^+(h_0, u_3u_1u_5')$. Since

$$\begin{aligned} f_\alpha(h_0, u_1u_2u_3u_4u_5) &= f^+(h_0, u_3u_1u_5u_4u_4u_3u_1) \\ &= f^+(f^+(h_0, u_3u_1u_5), u_4u_4u_3u_1) \\ &= f^+(f^+(h_0, u_3u_1u_5'), u_4u_4u_3u_1) \\ &= f_\alpha(h_0, u_1u_2u_3u_4u_5'), \end{aligned}$$

the set $\{u_1u_2u_3u_4u_5,\ u_1u_2u_3u_4u_5'\}$ is a nontrivial (2-)collision in f_α with the initial value h_0. Since, given $h \in \{0,1\}^n$ and $u \in \{0,1\}^{5m}$, the second block of

the message u is never used when calculating $f_\alpha(h,u)$, the set $\{u_1\, x\, u_3 u_4 u_5 \mid x \in \{0,1\}^m\}$ is a trivial 2^m-collision in f_α with any initial value.

Finally, we are ready to characterize a generalized iterated hash function. For each $j \in \mathbb{N}_+$, let $\alpha_j \in \mathbb{N}_j^+$ be such that $\mathrm{alph}(\alpha_j) = \mathbb{N}_j$. Denote $\hat{\alpha} = (\alpha_1, \alpha_2, \ldots)$. Define the *generalized iterated hash function* $H_{\hat{\alpha},f} : \{0,1\}^n \times (\{0,1\}^m)^+ \to \{0,1\}^n$ (based on $\hat{\alpha}$ and f) as follows: Given the initial value $h_0 \in \{0,1\}^m$ and the message x for which the block representation consists of j blocks, let $H_{\hat{\alpha},f}(h_0,x) = f_{\alpha_j}(h_0,x)$.

Remark 3.5. A traditional iterated hash function $H : (\{0,1\}^m)^+ \to \{0,1\}^n$ based on f (with initial value $h_0 \in \{0,1\}^n$) can of course be defined by $H(u) = f^+(h_0,u)$ for each $u \in (\{0,1\}^m)^+$. On the other hand H is a generalized iterated hash function $H_{\hat{\alpha},f} : \{0,1\}^n \times (\{0,1\}^m)^+ \to \{0,1\}^n$ based on $\hat{\alpha}$ and f where $\hat{\alpha} = (1, 1\cdot 2, 1\cdot 2\cdot 3, \ldots)$ and the initial value is fixed as h_0.

Now, let the generalized iterated hash function $H_{\hat{\alpha},f} : \{0,1\}^n \times (\{0,1\}^m)^+ \to \{0,1\}^n$ based on $\hat{\alpha}$ and f be as defined before the previous remark. Given $k \in \mathbb{N}_+$ and $h_0 \in \{0,1\}^m$, a *k-collision* in the generalized iterated hash function $H_{\hat{\alpha},f}$ is a set $A \subseteq (\{0,1\}^m)^+$ such that $|A| = k$ and for all $u, v \in A$, $|u| = |v|$ and $H_{\hat{\alpha},f}(h_0,u) = H_{\hat{\alpha},f}(h_0,v)$. Now suppose that A is a k-collision in $H_{\hat{\alpha},f}$ with initial value h_0. Let $l \in \mathbb{N}_+$ be such that $A \subseteq \{0,1\}^{ml}$, i.e., the length in blocks of each message in A is l. Then, by definition, for each $u, v \in A$, the equality $f_{\alpha_l}(h_0,u) = f_{\alpha_l}(h_0,v)$ holds. Since $\mathrm{alph}(\alpha_l) = \mathbb{N}_l$, the set A is a nontrivial k-collision in f_{α_l} with initial value h_0.

We assume that the attacker knows how $H_{\hat{\alpha},f}$ depends on the respective compression function f (i.e., the attacker knows $\hat{\alpha}$), but sees f only as a black box. She/he does not know anything about the internal structure of f and can only make *queries* (i.e., pairs $(h,x) \in \{0,1\}^n \times \{0,1\}^m$) on f and get the respective responses (values $f(h,x) \in \{0,1\}^n$).

A *k-collision attack on $H_{\hat{\alpha},f}$* is a probabilistic procedure (based on the birthday paradox) that finds a k-collision in $H_{\hat{\alpha},f}$ with probability equal to one for any initial value h_0. The *complexity of a k-collision attack on $H_{\hat{\alpha},f}$* is the expected number of queries on f required to get a k-collision.

3.2 Earlier work

In [10] Joux considers iterated hash functions $H : (\{0,1\}^m)^+ \to \{0,1\}^n$ and shows that, for each $r \in \mathbb{N}_+$ there exists a 2^r-collision attack on H of complexity $O(r \cdot 2^{n/2})$. The idea of Joux is simple and ingenious: a sequence of message sets

$\{u_{11}, u_{12}\}, \{u_{21}, u_{22}\}, \ldots, \{u_{r1}, u_{r2}\}$ such that $u_{i1} \neq u_{i2}$ and $f(h_{i-1}, u_{i1}) = f(h_{i-1}, u_{i2}) = h_i$ for $i = 1, 2, \ldots, r$ is generated with $O(r2^{n/2})$ queries on f. Then $H(y_1 y_2 \cdots y_r) = H(z_1 z_2 \cdots z_r)$ for all $y_1, z_1 \in \{u_{11}, u_{12}\}$, $y_2, z_2 \in \{u_{21}, u_{22}\}, \ldots, y_r, z_r \in \{u_{r1}, u_{r2}\}$, which means that the set $\{u_{11}, u_{21}\} \cdot \{u_{21}, u_{22}\} \cdots \{u_{r1}, u_{r2}\}$ is a 2^r-collision in H.

In [16] Nandi and Stinson show that there exists an attack procedure of complexity $O(r^2 \cdot (\ln r) \cdot (n + \ln(\ln 2r)) \cdot 2^{n/2})$ which (1) takes as input the (unique identity of) the function $f : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}^n$, a number $r \in \mathbb{N}_+$, and a word α such that alph(α) is sufficiently large and $|\alpha|_a \leq 2$ for all $a \in$ alph(α); (2) makes queries on f and gets the respective answers; and then (3) outputs (with probability equal to one) a 2^r-collision in f_α.

In [9] Hoch and Shamir continue the work on generalized iterated hash functions $H_{\hat{\alpha},f}$ showing the following: Let $\hat{\alpha} = (\alpha_1, \alpha_2, \ldots)$ and $q \in \mathbb{N}_+$ be such that $|\alpha_j|_i \leq q$ for all $j \in \mathbb{N}_+$, $i \in \mathbb{N}_j$. Then a polynomial $p(n, r)$ exists such that, for each $r \in \mathbb{N}_+$, a 2^r-collision attack on $H_{\hat{\alpha},f}$ of complexity $O(p(n, r)2^{\frac{n}{2}})$ can be constructed. However, some proofs are written in a short form and they contain a few inaccuracies being thus quite hard to follow. Our intention is to present new proofs and a detailed analysis of multicollisions in generalized iterated hash functions.

Multicollisions have been applied in practical attacks usually as a method for generating second preimages for hash values [13]. Multicollisions have been found for MD4, HAVAL, and Blender [13, 21]. The herding attack proposed by Kelsey and Kohno is also based on multicollisions [11]. Thus, theoretical advances in finding multicollisions have had an impact on the security of practical hash functions. However, there are no practical implementations of generalized iterated hash functions and thus theoretical advances in this field have a limited influence in practice, but can help in devising more secure hash functions in the future.

3.3 Nested Multicollision Attack Schema (*NMCAS*)

Below we describe a general (and at this stage still informal) attack procedure that, given $H_{\hat{\alpha},f}$, $h_0 \in \{0,1\}^n$, and $r \in \mathbb{N}_+$ creates a 2^r-collision in the generalized iterated hash function $H_{\hat{\alpha},f}$ with initial value h_0.

Procedure Schema *NMCAS*

Input: A generalized iterated hash function $H_{\hat{\alpha},f}$, an initial value $h_0 \in \{0,1\}^n$, a positive integer r.
Output: A 2^r-collision in $H_{\hat{\alpha},f}$.
Step 1: Choose (a large) $l \in \mathbb{N}_+$. Consider the lth element α_l of the sequence $\hat{\alpha}$. Let $\alpha_l = i_1 i_2 \cdots i_s$, where $s \in \mathbb{N}_+$ and $i_j \in \mathbb{N}_l$ for $j = 1, 2, \ldots, s$.

Step 2: Fix a (large) set of *active indices* $\text{Act} \subseteq \mathbb{N}_l = \{1, 2, \ldots, l\}$.
Step 3: Factorize the word α_l into nonempty strings appropriately, i.e., find $p \in \{1, 2, \ldots, s\}$ and $\beta_i \in \mathbb{N}_l^+$ such that $\alpha_l = \beta_1\beta_2\cdots\beta_p$.
Step 4: Based upon the active indices, create a large multicollision in f_{β_1}. More exactly, find message block sets $M_1, M_2, \ldots, M_l$ satisfying the following properties.

(i) If $i \in \mathbb{N}_l \setminus \text{Act}$, then the set M_i consists of one constant message block ω.

(ii) If $i \in \text{Act}$, then the set M_i consists of two different message blocks m_{i1} and m_{i2}.

(iii) The set $M = M_1M_2\cdots M_l = \{u_1u_2\cdots u_l \,|\, u_i \in M_i,\ i = 1, 2, \ldots, l\}$ is a $2^{|\text{Act}|}$-collision in f_{β_1} with initial value h_0.

Step 5: Based on the set $C_1 = M$, find message sets $C_2, C_3, \ldots, C_p$ such that

(iv) $C_p \subseteq C_{p-1} \subseteq \cdots \subseteq C_1 = M$.

(v) For each $j \in \{1, 2, \ldots, p\}$ the set C_j is a (large) multicollision in $f_{\beta_1\beta_2\cdots\beta_j}$ with initial value h_0.

(vi) $|C_p| = 2^r$.

Step 6: Output C_p.

It should be clear that if the above procedure is successfully carried out, then

$$H_{\hat{\alpha},f}(h_0, m) = H_{\hat{\alpha},f}(h_0, m')$$

for all $m, m' \in C_p$. Also one should note that *NMCAS* can be applied trivially to produce a 2^r-collision with initial value h_0 for any generalized iterated hash function $H_{\hat{\alpha},f}$. Namely, choosing $l \geq n + r$ we know, by the pigeonhole principle, that among the messages of length l, a 2^r-collision exists. Then letting $Act = \mathbb{N}_l$ and $p = 1$, we can, by going in the worst case through all the 2^{n+r} possible message values, certainly find the desired multicollision. Note that, as proved in [18], by hashing $((2^r)!)^{\frac{1}{2^r}} 2^{\frac{n(2^r-1)}{2^r}}$ messages, a 2^r-collision is found with probability approx. 0.5.

Given $H_{\hat{\alpha},f}$ and r, does there exist a 2^r-collision attack on $H_{\hat{\alpha},f}$ of complexity $O(2^{\frac{n}{2}})$? The problem in its full generality, i.e., with no restrictions on $H_{\hat{\alpha},f}$, seems to be extremely difficult and is certainly still open. Probably the answer to the question is negative.

Call the sequence $\hat{\alpha} = (\alpha_1, \alpha_2, \ldots)$ *q-bounded*, $q \in \mathbb{N}_+$, if $|\alpha_j|_i \leq q$ for each $j \in \mathbb{N}_+$ and $i \in \mathbb{N}_j$. Now suppose that $q \in \mathbb{N}_+$ and in $H_{\hat{\alpha},f}$ the sequence $\hat{\alpha}$ is q-bounded. In the following we shall show that the procedure *NMCAS* with input

$H_{\hat{\alpha},f}$, h_0 and r can be realized so that a 2^r-collision in $H_{\hat{\alpha},f}$ with initial value h_0 is created (with probability equal to one) and the expected number of queries on the compression function f is $O(\tilde{p}(n,r)\,2^{\frac{n}{2}})$ where $\tilde{p}(n,r)$ is a polynomial.

The idea behind the successful construction is the fact that since $\hat{\alpha}$ is q-bounded, unavoidable regularities start to appear in the word α_l of $\hat{\alpha}$ when l is increased. More accurately, choosing l big enough (still so that $|\alpha_l|$ depends only polynomially on n and r), arbitrarily large sets $A \subseteq \mathrm{alph}(\alpha_l)$ can be found such that

(P1) $\alpha_l = \beta_1\beta_2\cdots\beta_p$, where $p \in \{1,2,\ldots,q\}$, β_i is a word such that $A \subseteq \mathrm{alph}(\beta_i)$ and the elements of A are independent with respect to $\prec_{\beta_i}$ for $i = 1,2,\ldots,p$; and

(P2) for any $i \in \{1,2,\ldots,p-1\}$, if $\pi_A(\beta_i) = z_1 z_2 \cdots z_{n^{p-i}k}$ is a factorization of $\pi_A(\beta_i)$ such that $|\mathrm{alph}(z_j)| = n^{i-1}$ for $j = 1,2,\ldots,n^{p-i}k$ and $\pi_A(\beta_{i+1}) = u_1 u_2 \cdots u_{n^{p-i-1}k}$ is a factorization of $\pi_A(\beta_{i+1})$ such that $|\mathrm{alph}(u_j)| = n^i$ for $j = 1,2,\ldots,n^{p-i-1}k$, then for each $j_1 \in \{1,2,\ldots,n^{p-i}k\}$, there exists $j_2 \in \{1,2,\ldots,n^{p-i-1}k\}$ such that $\mathrm{alph}(z_{j_1}) \subseteq \mathrm{alph}(u_{j_2})$.

The property (P1) allows us to construct a $2^{|A|}$-collision C_1 in f_{β_1} with any initial value h_0 so that the expected number of queries on f is $O(|\beta_1|2^{\frac{n}{2}})$. The property (P2) ensures that based on the multicollision guaranteed by (P1), we can proceed and create the multicollision C_i in $f_{\beta_1\beta_2\cdots\beta_i}$ so that (i) the expected number of queries on f is $O(|\beta_1\beta_2\cdots\beta_i|2^{\frac{n}{2}})$ for all $i = 2,3,\ldots,p$; and (ii) the cardinality of C_p is 2^r. Since $|\alpha_l|$ (and thus $|\beta_1\beta_2\cdots\beta_i|$ for $i = 1,2,\ldots,p$) depend only polynomially on n and r, steps 1 to 6 in *NMCAS* do not consume too much resources.

We prove the necessary combinatorial results for properties (P1) and (P2) in the next section. The construction of the actual attack is postponed to Section 5.

Remark 3.6. In many problems of combinatorics on words (in contrast to ours), arbitrarily long words over a fixed (finite) alphabet are considered. Then, as the length of the word increases, unavoidable regularities start to appear and some famous results of classical combinatorics like Ramsey's, Shirshov's and Van der Waerden's Theorems may be applied (for details, see for instance the book of de Luca and Varricchio [5]).

4 Basic combinatorial results

Let α be a (nonempty) word and A any alphabet. We wish to study how the occurrences in α of any symbol $a \in A$ are positioned in relation to occurrences in α of other symbols of A. In principle, for this purpose the image $\pi_A(\alpha)$ of α

under the projection morphism π_A is completely sufficient. However, for the sake of simpler notation and ability to apply some classical results of combinatorics directly, one more concept is introduced (see also [9, 16]): define $(\alpha)_A = \epsilon$ if $\pi_A(\alpha) = \epsilon$ and $(\alpha)_A = a_1 a_2 \cdots a_s$ if $\pi_A(\alpha) \in a_1^+ a_2^+ \cdots a_s^+$, where $s \in \mathbb{N}_+$, $a_1, a_2, \ldots, a_s \in A$, and $a_i \neq a_{i+1}$ for $i = 1, 2, \ldots, s-1$.

It should be obvious from the definition that the word α_A exists and is unique.

Example 4.1. Let $\alpha = a_5 a_2^3 a_7^2 a_5^2 a_1 a_4^5 a_3^2 a_1^2 a_2 a_6^3$, so $\mathrm{alph}(\alpha) = \{a_1, a_2, \ldots, a_7\}$. Let us choose $A = \mathrm{alph}(\alpha) \setminus \{a_5\}$. Then $\pi_A(\alpha) = a_2^3 a_7^2 a_1 a_4^5 a_3^2 a_1^2 a_2 a_6^3$ and $(\alpha)_A = a_2 a_7 a_1 a_4 a_3 a_1 a_2 a_6$.

Note that even though the word α_A is unique, there certainly may be different ways to obtain it from the original word α. For example, let $\alpha = abbcc$ and $A = \{b, c\}$. Now, $\alpha_A = bc$, but there are four different ways of obtaining this word from α depending on whether one chooses the first or the second occurrence of b and c.

Remark 4.2. It is important to notice that the operation $(\cdot)_A$ does not behave like a morphism, i.e., given an alphabet A and words u, v, we generally have $(uv)_A \neq u_A v_A$. The case where $A = \{a\}$ and $(aa)_A = a \neq aa = (a)_A (a)_A$ is the simplest possible example. This certainly means that in all cases $(uv)_A$ cannot be constructed from u_A and v_A (as is done in [9]).

Remark 4.3. Let α be a word, $\alpha \neq \epsilon$, and $A \subseteq \mathrm{alph}(\alpha)$ nonempty. Recall that the independence of elements in A with respect to $\prec_\alpha$ means that these elements form a chain in the partially ordered set $(A, \prec_\alpha)$. Then the following conditions are equivalent.

(a) The elements of A are independent with respect to $\prec_\alpha$.

(b) There exists a sequence $a_1, a_2, \ldots, a_d$ of all $d = |A|$ elements of A such that $\pi_A(\alpha)$ is in $a_1^+ a_2^+ \cdots a_d^+$.

(c) The word α_A is a permutation of A.

Suppose that $\hat{\alpha} = (\alpha_1, \alpha_2, \ldots)$ is q-bounded, $q \in \mathbb{N}_+$, i.e., for each $j \in \mathbb{N}_+$ and $i \in \mathbb{N}_j$, the inequality $|\alpha_j|_i \leq q$ is satisfied. Our first task is to show that the property (P1) holds.

We state the following (binary) matrix form of Hall's famous matching theorem (see, for instance [2, p. 77]).

Theorem 4.4 (Hall). *Let m and n, $m \leq n$, be positive integers and $A = (a_{ij})_{m \times n}$ be a $m \times n$-dimensional binary matrix. Now there exists an injective function*

$\sigma : \{1, 2, \ldots, m\} \to \{1, 2, \ldots, n\}$ such that $a_{i\sigma(i)} = 1$ for $i = 1, 2, \ldots, m$ if and only if for each $I \subseteq \{1, 2, \ldots, m\}$ the number of elements a_{ij} such that $i \in I$, $j \in \{1, 2, \ldots, n\}$ and $a_{ij} = 1$ is at least $|I|$.

It is a well-known fact that Dilworth's Theorem and Hall's Theorem are, as many results in basic combinatorics, strongly related. The following theorem can be found also in [9].

Theorem 4.5 (Partition Theorem). *Let $k \in \mathbb{N}_+$ and A be a finite nonempty set such that k divides $|A|$. Furthermore, let $\{B_i\}_{i=1}^k$ and $\{C_j\}_{j=1}^k$ be partitions of A such that $|B_i| = |C_j|$ for $i, j = 1, 2, \ldots, k$. Then for each $x \in \mathbb{N}_+$ such that $|A| \geq k^3 \cdot x$, there exists a bijection $\sigma : \{1, 2, \ldots, k\} \to \{1, 2, \ldots, k\}$ for which $|B_i \cap C_{\sigma(i)}| \geq x$ for $i = 1, 2, \ldots, k$.*

Proof. Let $x \in \mathbb{N}_+$ be such that $|A| \geq k^3 \cdot x$. Let $D = (d_{ij})_{k\times k}$ be the $k \times k$-dimensional binary matrix defined by $d_{ij} = 1$ if $|B_i \cap C_j| \geq x$ and $d_{ij} = 0$ if $|B_i \cap C_j| < x$ for each $i, j \in \{1, 2, \ldots, k\}$.

Suppose that the required bijection does not exist. By Hall's Theorem, we can find a set $I \subseteq \{1, 2, \ldots, n\}$ of size $r \leq n$ such that the number s of elements d_{ij} for which $i \in I$, $j \in \{1, 2, \ldots, n\}$, and $d_{ij} = 1$ is less than r. Assume without loss of generality, that $d_{ij} = 0$ for $i = 1, 2, \ldots, r$, $j = s + 1, s + 2, \ldots, n$. This means that $|B_i \cap C_j| < x$ for each $i \in \{1, 2, \ldots, r\}$, $j \in \{s + 1, s + 2, \ldots, n\}$. Certainly $\sum_{i=1}^s |C_i| = s \cdot \frac{|A|}{k}$. On the other hand (since $|B_i \cap C_j| < x$ for each $i \in \{1, 2, \ldots, r\}$, $j \in \{s + 1, s + 2, \ldots, n\}$) we have

$$\sum_{i=1}^{s} |C_i| \geq r \cdot \frac{|A|}{k} - r(k - s)(x - 1).$$

Then $s \cdot \frac{|A|}{k} \geq r \cdot \frac{|A|}{k} - r(k-s)(x-1)$, i.e., $(r-s) \cdot \frac{|A|}{k} - r(k-s)(x-1) \leq 0$. We have reached a contradiction, since $(r-s) \cdot \frac{|A|}{k} - r(k-s)(x-1) \geq (r-s)k^2x - r(k-s)(x-1) > 0$. □

Remark 4.6. In the previous theorem the power 3 of k cannot be reduced to 2. Consider the following example. Let A be a set consisting of $k^2 \cdot x$ elements, where $k, x \in \mathbb{N}_+$, $x \geq k^2$. Suppose $r \in \{1, 2, \ldots, k - 2\}$ and let $\{A_i\}_{i=1}^k$ and $\{B_i\}_{i=1}^k$ be two partitions of A such that $|A_i \cap B_1| = x + k - r$ for $i = 1, 2, \ldots, r + 1$; $|A_i \cap B_j| = x$ for $i = 1, 2, \ldots, r + 1$, $j = 2, 3, \ldots, r$; $|A_i \cap B_j| = x - 1$ for $i = 1, 2, \ldots, r + 1$, $j = r + 1, r + 2, \ldots, k$; $|A_{r+2} \cap B_1| = x - (r + 1)(k - r)$; $|A_{r+2} \cap B_j| = x$ for $j = 2, 3, \ldots, r$; $|A_{r+2} \cap B_j| = x + r + 1$ for $j = r + 1$, $r + 2, \ldots, k$; and $|A_i \cap B_j| = x$ for $i = r + 3, r + 4, \ldots, k$, $j = 1, 2, \ldots, k$. Then

$|A| = k^2 \cdot x$ and $|A_i| = |B_i| = k \cdot x$ for $i = 1, 2, \ldots, k$. Clearly there does not exist a bijection σ of $\{1, 2, \ldots, k\}$ such that $|A_i \cap B_{\sigma(i)}| \geq x$ for $i = 1, 2, \ldots, k$.

The example above generalizes neither to the case $|A| \geq a \cdot k^2$ where $a \geq 2$ nor to the case $|A| \geq k^b$ where b is a rational number such that $2 < b < 3$.

Remark 4.7. Let A be a finite set, $k \in \mathbb{N}_+$, and $\{A_i\}_{i=1}^k$ and $\{B_i\}_{i=1}^k$ two partitions of A such that $|A_i| = |B_j|$ for all $i, j \in \{1, 2, \ldots, k\}$. When applying Lemma 4.5, the total number $\sum_{i=1}^k |A_i \cap B_{\sigma(i)}|$ of elements in the intersections can be guaranteed to be at least $\frac{|A|}{k^2}$. This implies that in Theorem 1 of [9] one has to assume that $l = |M| \geq k^{2q-3} \cdot n^{(q-1)^2}$ instead of $l = |M| \geq k^3 \cdot n^{3(q-3)+2}$ (see the proof of Theorem 4.15). Note that the assumption leads to a remarkable increase in the complexity of the respective multicollision attack presented later.

The following lemma is a new formulation of a result in [16].

Lemma 4.8. *Let m, n and q be positive integers and α a word such that* $\text{alph}(\alpha) \geq m \cdot n$. *Then either* (i) *the maximum chain length of* $(\text{alph}(\alpha), \prec_\alpha)$ *is at least m; or* (ii) *the maximum number of pairwise incomparable elements in* $(\text{alph}(\alpha), \prec_\alpha)$ *is greater than n.*

Proof. Suppose that the maximum chain length in $(\text{alph}(\alpha), \prec_\alpha)$, denoted by d, is less than m. Let t be the minimum number of chains needed to cover $(\text{alph}(\alpha), \prec_\alpha)$. Obviously

$$m \cdot n \leq |\,\text{alph}(\alpha)| \leq d \cdot t.$$

Since $d < m$, we have $t > n$. By Dilworth's Theorem, the maximum number of pairwise incomparable elements of $(\text{alph}(\alpha), \prec_\alpha)$ is equal to t. □

Remark 4.9. Note that the limits given by the previous lemma are sharp in the sense that for each $m, n \in \mathbb{N}_+$, there exists a word α such that $|\text{alph}(\alpha)| = m \cdot n$, the maximum chain length in $(\text{alph}(\alpha), \prec_\alpha)$ is equal to m and the maximum number of pairwise incomparable elements in the set $(\text{alph}(\alpha), \prec_\alpha)$ is equal to n. Clearly the word $(a_{11}a_{12}\cdots a_{1n})^2(a_{21}a_{22}\cdots a_{2n})^2\cdots(a_{m1}a_{m2}\cdots a_{mn})^2$ is an example of such an α.

Theorem 4.10. *For all positive integers q and m there exist positive integers r_q and s_q with the following property. Let α be a word such that* $|\text{alph}(\alpha)| \geq r_q \cdot m^{s_q}$ *and* $|\alpha|_a \leq q$ *for each* $a \in \text{alph}(\alpha)$. *Then there exists* $A \subseteq \text{alph}(\alpha)$ *with* $|A| \geq m$ *and* $p \in \{1, 2, \ldots, q\}$ *as well as words* $\alpha_1, \alpha_2, \ldots, \alpha_p$ *such that* $\alpha = \alpha_1\alpha_2\cdots\alpha_p$ *and for all* $i \in \{1, 2, \ldots, p\}$, *the word* $(\alpha_i)_A$ *is a permutation of A.*

Proof. Proceed by induction on q.

Let $q = 1$. Choose $r_1 = s_1 = 1$ and $B = \mathrm{alph}(\alpha)$. Then α is a permutation of all the letters in $\mathrm{alph}(\alpha)$ and the claim is satisfied.

Assume that positive integers $r_1, s_1, r_2, s_2, \ldots, r_{q-1}, s_{q-1}$ such that $r_{i-1} < r_i$ and $s_{i-1} < s_i$ for $i = 2, 3, \ldots, q-1$ satisfying the claim of the lemma have been determined.

Let $r_q = q^{s_{q-1}} r_{q-1}^{s_{q-1}+1}$ and $s_q = s_{q-1}^2 + 1$. Suppose further that α is a word such that $\mathrm{alph}(\alpha) \geq r_q \cdot m^{s_q}$ and $|\alpha|_a \leq q$ for all $a \in \mathrm{alph}(\alpha)$. Let c be the maximum chain length and d the maximum number of pairwise incomparable elements in $(\mathrm{alph}(\alpha), \prec_\alpha)$. By the previous lemma, either $c \geq m$ or $d \geq q^{s_{q-1}} r_{q-1}^{s_{q-1}+1} m^{s_{q-1}^2}$.

In the former case, let V be any largest possible set of pairwise incomparable elements in $(\mathrm{alph}(\alpha), \prec_\alpha)$. Certainly $|V| \geq m$; choosing $A = V$ and $p = 1$ (thus $\alpha_1 = \alpha$) we find that the induction is extended.

Consider the latter case and assume that $d \geq q^{s_{q-1}} r_{q-1}^{s_{q-1}+1} m^{s_{q-1}^2}$. Let U be a set of pairwise incomparable elements in $(\mathrm{alph}(\alpha), \prec_\alpha)$ such that $|U| = d$. Let $u \in U$ and $\alpha', \alpha'' \in \mathrm{alph}(\alpha)^+$ be such that $\alpha = \alpha' u \alpha''$, $|\alpha'|_u = 0$ and for all $x \in U$, $x \neq u$, $|\alpha'|_x > 0$. Since the elements of U are pairwise incomparable, we have $0 < |\alpha' u|_z \leq q-1$ and $|\alpha''|_z \leq q-1$ for all $z \in U$. Let $\beta = \pi_U(\alpha)$, where π_U is the projection morphism: $\mathrm{alph}(\alpha)^* \to U^*$. Then $\beta = \beta'\beta''$ where $\beta' = \pi_U(\alpha' u)$ and $\beta'' = \pi_U(\alpha'')$. By the facts above, $\mathrm{alph}(\beta) = \mathrm{alph}(\beta') = U$, $U \setminus \{u\} \subseteq \mathrm{alph}(\beta'')$, $|\beta'|_x \leq q-1$, and $|\beta''|_x \leq q-1$ for each $x \in U$.

Apply the induction hypothesis on β'. Since $|\mathrm{alph}(\beta')| \geq q^{s_{q-1}} r_{q-1}^{s_{q-1}+1} m^{s_{q-1}^2}$ and

$$q^{s_{q-1}} r_{q-1}^{s_{q-1}+1} m^{s_{q-1}^2} \;\geq\; r_{q-1}(1 + r_1 m^{s_1} + \cdots + r_{q-1} m^{s_{q-1}})^{s_{q-1}},$$

there exists an alphabet $B \subseteq \mathrm{alph}(\beta)$, $|B| \geq 1 + r_1 m^{s_1} + \cdots + r_{q-1} m^{s_{q-1}}$, $k_1 \in \{1, 2, \ldots, q-1\}$, and words $\beta_1, \beta_2, \ldots, \beta_{k_1}$ such that $\beta' = \beta_1 \beta_2 \cdots \beta_{k_1}$ and for each $i \in \{1, 2, \ldots, k_1\}$, $b \in B$, we have $|(\beta_i)_B|_b = 1$. Now consider the word β''. Remember that $|\beta''|_b \leq q-1$ for all $b \in B$. For each $i \in \{0, 1, \ldots, q-1\}$, let $B_i = \{b \in B \mid |\beta''|_b = i\}$. Certainly the sets $B_1, B_2, \ldots, B_{q-1}$ are pairwise disjoint and $B = \bigcup_{i=0}^{q-1} B_i$. Furthermore, either $B_0 = \emptyset$ or $B_0 = \{u\}$. Since $|B| \geq 1 + r_1 m^{s_1} + \cdots + r_{q-1} m^{s_{q-1}}$, there exists, by the pigeonhole principle, an integer $i \in \{1, 2, \ldots, q-1\}$ such that $|B_i| \geq r_i m^{s_i}$. Let i_0 be such an i. Consider the word $\gamma = \pi_{B_{i_0}}(\beta'')$ where $\pi_{B_{i_0}}$ is the projection morphism : $\mathrm{alph}(\alpha)^* \to B_{i_0}^*$. Again, by the induction hypothesis, there exists $C \subseteq B_{i_0}$, $|C| \geq m$ and $k_2 \in \{1, 2, \ldots, q-1\}$, and words $\gamma_1, \gamma_2, \ldots, \gamma_{k_2}$ such that $\gamma = \gamma_1 \gamma_2 \cdots \gamma_{k_2}$ and for each $i \in \{1, 2, \ldots, k_2\}$, $c \in C$, we have $|(\gamma_i)_C|_c = 1$. Since $C \subseteq B$, we have $|(\beta_i)_C|_c = 1$ for each $i \in \{1, 2, \ldots, k_2\}$ and $c \in C$. Certainly $k_1 + k_2 \leq$

q. Choose $A = C$, $p = k_1 + k_2$, and words $\alpha_1, \alpha_2, \ldots, \alpha_{k_1+k_2}$ so that $\alpha = \alpha_1\alpha_2\cdots\alpha_{k_1+k_2}$, $\pi_C(\alpha_i) = \pi_C(\beta_i)$ for $i = 1, 2, \ldots, k_1$, and $\pi_C(\alpha_{k_1+i}) = \gamma_i$ for $i = 1, 2, \ldots, k_2$ where π_C is the projection morphism : $\mathrm{alph}(\alpha)^* \to C^*$. The proof is now complete. □

Remark 4.11. The first task just after the induction hypothesis in the proof above is to catch the chain of the maximum length in the set $(\mathrm{alph}(\alpha), \prec_\alpha)$. If this entity is at least m, we are certainly done. If not, we concentrate on incomparable elements in $(\mathrm{alph}(\alpha), \prec_\alpha)$ and erase everything else. This is quite natural, since while wishing to have more than one permutation in our subword construction, only incomparable elements can be brought into the play. Our final task is to find an alphabet $A \subseteq \mathrm{alph}(\alpha)$ and a decomposition $\alpha = \alpha_1\alpha_2\cdots\alpha_p$, $p \leq q$, such that, for each $i \in \{1, 2, \ldots, p\}$, the elements of A form a chain in $(\mathrm{alph}(\alpha_i), \prec_{\alpha_i})$. This is equivalent to saying that, for each $i \in \{1, 2, \ldots, p\}$, the word $(\alpha_i)_A$ is a permutation of A.

Remark 4.12. The parameters r_q and s_q in Theorem 4.10 grow very fast with respect to q, the parameter restricting the number of occurrences of any symbol in α. For s_q we have the recurrence relations

$$\begin{cases} s_1 = 1 \\ s_{q+1} = s_q^2 + 1, \quad \text{if } q \in \mathbb{N}_+. \end{cases}$$

We can roughly estimate that $s_q^2 < s_{q+1} < 2s_q^2$ for each $q \in \mathbb{N}$. It is easily seen that s_q is in $\Omega(2^{2^{q-1}})$ and in $O(2^{2^q-1})$. On the other hand

$$\begin{cases} r_1 = 1 \\ r_{q+1} = (q+1)^{s_q} r_q^{s_q+1}, \quad \text{if } q \in \mathbb{N}_+. \end{cases}$$

Again, with a rough estimate, $(q+1)^{s_q} r_q^{s_q} < r_{q+1} < r_q^{2s_q}$ for all $q \in \mathbb{N}$, $q \geq 2$. With a standard consideration we find that r_q is in $\Omega(2^{2^{2^{q-1}-1}})$ and in $O(2^{2^{2^q-3}})$. This, among other things, limits the appliance of the lemma substantially. It means that one can apply the lemma only to those words where $\mathrm{alph}(\alpha)$ is very large when compared with q. The sequences of numbers generated by recursions that are similar to s_q and r_q have been studied for example in [1].

Recall that the infinite sequence $\hat{\alpha} = (\alpha_1, \alpha_2, \ldots)$ of words is such that for all $l \in \mathbb{N}_+$, we have $\mathrm{alph}(\alpha_l) = \mathbb{N}_l$, i.e., α_l is a word over the alphabet $\mathbb{N}_l = \{1, 2, \ldots, l\}$, and each symbol of $\mathbb{N}_l$ occurs in α_l. For any (probabilistic) algorithm to be able to use $\hat{\alpha}$, the sequence has to be *effectively encoded*, i.e., it has to

have a finite presentation from which the word α_i can be computed in polynomial time with respect to $|\alpha_i|$ for all $i \in \mathbb{N}_+$.

We wish to remind that a permutation of an alphabet A is any word $w \in A^+$ such that $|w|_a = 1$ for each $a \in A$.

The result we achieved in our previous theorem is not yet sufficient for our purposes; inside the permutations $(\alpha_1)_A, (\alpha_2)_A, \ldots, (\alpha_k)_A$ of A, the symbols have to be appropriately grouped. We need an application of the following lemma.

Lemma 4.13. *Let $d_0, d_1, d_2, \ldots, d_r$, where $r \in \mathbb{N}_+$ be positive integers such that d_i divides d_{i-1} for $i = 1, 2, \ldots, r$, A an alphabet of cardinality $|A| = d_0 d_1^2 d_2^2 \cdots d_r^2$, and $w_1, w_2, \ldots, w_{r+1}$ permutations of A. Then there exists a subset B of A of cardinality $|B| = d_0$ such that the following conditions are satisfied.*

(1) *For any $i \in \{1, 2, \ldots, r\}$, if $\pi_B(w_i) = x_1 x_2 \cdots x_{d_i}$ is the factorization of $\pi_B(w_i)$ and $\pi_B(w_{i+1}) = y_1 y_2 \cdots y_{d_i}$ is the factorization of $\pi_B(w_{i+1})$ into d_i equal length $(= \frac{d_0}{d_i})$ blocks, then for each $j \in \{1, 2, \ldots, d_i\}$, there exists $j' \in \{1, 2, \ldots, d_i\}$ such that $\mathrm{alph}(x_j) = \mathrm{alph}(y_{j'})$; and*

(2) *If $w_r = z_1 z_2 \cdots z_{d_r}$ and $w_{r+1} = u_1 u_2 \cdots u_{d_r}$ are factorizations of w_r and w_{r+1}, respectively, into d_r equal length $(= d_0 d_1^2 d_2^2 \cdots d_{r-1}^2 d_r)$ blocks, then the words*

$$\pi_B(w_r) = \pi_B(z_1)\pi_B(z_2)\cdots\pi_B(z_{d_r}) \quad \text{and}$$
$$\pi_B(w_{r+1}) = \pi_B(u_1)\pi_B(u_2)\cdots\pi_B(u_{d_r})$$

are factorizations of $\pi_B(w_r)$ and $\pi_B(w_{r+1})$, respectively, into d_r equal length $(= \frac{d_0}{d_r})$ blocks.

Proof. Proceed by induction on r. Consider first the case $r = 1$. Let $w_1 = z_1 z_2 \cdots z_{d_1}$ and $w_2 = u_1 u_2 \cdots u_{d_1}$ be factorizations of w_1 and w_2, respectively, into d_1 equal length $(= d_0 d_1)$ blocks. Then $\{\mathrm{alph}(z_i)\}_{i=1}^{d_1}$ and $\{\mathrm{alph}(u_i)\}_{i=1}^{d_1}$ are partitions of A into equal cardinality $(= d_0 d_1)$ sets. Now $|A| = \frac{d_0}{d_1} d_1^3$, so by the Partition Theorem, there exists a bijection from $\sigma : \{1, 2, \ldots, d_1\}$ onto $\{1, 2, \ldots, d_1\}$ such that $|\{\mathrm{alph}(z_i)\} \cap \{\mathrm{alph}(u_{\sigma(i)})\}| \geq \frac{d_0}{d_1}$ for $i = 1, 2, \ldots d_1$. Let σ be as above and $B_i \subseteq \{\mathrm{alph}(z_i)\} \cap \{\mathrm{alph}(u_{\sigma(i)})\}$ such that $|B_i| = \frac{d_0}{d_1}$ for $i = 1, 2, \ldots, d_1$. Denote $B = \bigcup_{i=1}^{d_1} B_i$. Then certainly $|B| = d_0$ and $\{B_i\}_{i=1}^{d_1}$ is a partition of B. Define $x_i = \pi_B(z_i)$ and $y_i = \pi_B(u_i)$ for $i = 1, 2, \ldots, d_1$. Then

$$\pi_B(w_1) = \pi_B(z_1)\pi_B(z_2)\cdots\pi_B(z_{d_1}) = x_1 x_2 \cdots x_{d_1}; \quad \text{and}$$
$$\pi_B(w_2) = \pi_B(u_1)\pi_B(u_2)\cdots\pi_B(u_{d_1}) = y_1 y_2 \cdots y_{d_1}$$

where $|x_i| = |y_i| = \frac{d_0}{d_1}$ and $\text{alph}(x_i) = \text{alph}(y_{\sigma(i)}) = B_i$ for $i = 1, 2, \ldots, d_1$. Thus (1) and (2) hold for $r = 1$.

Now suppose that the lemma holds for $r = k$, $k \in \mathbb{N}_+$. Consider the case $r = k + 1$. Let $w_{k+1} = z'_1 z'_2 \cdots z'_{d_{k+1}}$ and $w_{k+2} = u'_1 u'_2 \cdots u'_{d_{k+1}}$ be factorizations of w_{k+1} and w_{k+2}, respectively, into d_{k+1} equal length $(= d_0 d_1^2 d_2^2 \cdots d_k^2 d_{k+1})$ blocks. Then again $\{\text{alph}(z'_i)\}_{i=1}^{d_{k+1}}$ and $\{\text{alph}(u'_i)\}_{i=1}^{d_{k+1}}$ are partitions of A into equal cardinality $(= d_0 d_1^2 d_2^2 \cdots d_k^2 d_{k+1})$ sets. Again

$$|A| = \frac{d_0 d_1^2 d_2^2 \cdots d_k^2}{d_{k+1}} d_{k+1}^3$$

so, by the Partition Theorem, there exists a bijection $\sigma : \{1, 2, \ldots, d_{k+1}\} \to \{1, 2, \ldots, d_{k+1}\}$, with $|\{\text{alph}(z'_i)\} \cap \{\text{alph}(u'_{\sigma(i)})\}| \geq \frac{d_0 d_1^2 d_2^2 \cdots d_k^2}{d_{k+1}}$ for $i = 1, 2, \ldots, d_{k+1}$. Let σ be as above and $d' = \frac{d_0 d_1^2 d_2^2 \cdots d_k^2}{d_{k+1}}$. Let $C_i \subseteq \{\text{alph}(z'_i)\} \cap \{\text{alph}(u'_{\sigma(i)})\}$ such that $|C_i| = d'$ for $i = 1, 2, \ldots, d_{k+1}$. Denote $C = \bigcup_{i=1}^{d_{k+1}} C_i$. Then $|C| = d' d_{k+1} = d_0 d_1^2 d_2^2 \cdots d_k^2$ and $\{C_i\}_{i=1}^{d_{k+1}}$ is a partition of C. Define $w'_i = \pi_C(w_i)$ for $i = 1, 2, \ldots, k + 2$. Obviously $w'_{k+1} = \pi_C(z'_1)\pi_C(z'_2) \cdots \pi_C(z'_{d_{k+1}})$ and $w'_{k+2} = \pi_C(u'_1)\pi_C(u'_2) \cdots \pi_C(u'_{d_{k+1}})$ are factorizations of w'_{k+1} and w'_{k+2}, respectively, into d_{k+1} equal length $(= d')$ blocks such that $|\text{alph}(\pi_C(z'_i))| = |\text{alph}(\pi_C(u'_i))| = d'$ and $\text{alph}(\pi_C(z'_i)) = \text{alph}(\pi_C(u'_{\sigma(i)}))$ for $i = 1, 2, \ldots, d_{k+1}$.

Certainly C is an alphabet of cardinality $|C| = d_0 d_1^2 d_2^2 \cdots d_k^2$ and $w'_1, w'_2, \ldots, w'_{k+1}$ (as well as w'_{k+2}) are permutations of C. Apply the induction hypothesis to achieve an alphabet $B \subseteq C \subseteq A$ so that (1) and (2) hold when r is replaced by k and w_i is replaced by w'_i for $i = 1, 2, \ldots, k$. Without loss of generality we may interpret π_B to be the projection morphism: $A^* \to B^*$. Since $\pi_B(w_i) = \pi_B(w'_i)$ for $i = 1, 2, \ldots, k + 2$, the condition (1) holds for $i = 1, 2, \ldots, k$. Consider the case $i = k + 1$. Let $w'_{k+1} = z_1 z_2 \cdots z_{d_k}$ be the factorization of w'_{k+1} into d_k equal length $(= d_0 d_1^2 d_2^2 \cdots d_{k-1}^2 d_k)$ blocks. Then, by property (2) of the induction hypothesis, $\pi_B(w'_{k+1}) = \pi_B(z_1)\pi_B(z_2) \cdots \pi_B(z_{d_k})$ is the factorization of $\pi_B(w'_{k+1})$ into d_k equal length $(= \frac{d_0}{d_k})$ blocks. By our earlier considerations, $\pi_C(w'_{k+1}) = \pi_C(z'_1)\pi_C(z'_2) \cdots \pi_C(z'_{d_{k+1}})$ is a factorization of $\pi_C(w'_{k+1})$ into d_{k+1} equal length $(= d' = \frac{d_0 d_1^2 d_2^2 \cdots d_k^2}{d_{k+1}})$ blocks. Since d_{k+1} divides d_k, the equality

$$\pi_C(z'_i) = z_{(i-1)\frac{d_k}{d_{k+1}}+1} z_{(i-1)\frac{d_k}{d_{k+1}}+2} \cdots z_{i\frac{d_k}{d_{k+1}}}$$

holds for $i = 1, 2, \ldots, d_{k+1}$. Then

$$\pi_B(z'_i) = \pi_B\big(z_{(i-1)\frac{d_k}{d_{k+1}}+1}\big)\pi_B\big(z_{(i-1)\frac{d_k}{d_{k+1}}+2}\big)\cdots\pi_B\big(z_{i\frac{d_k}{d_{k+1}}}\big)$$

and $|\mathrm{alph}(\pi_B(z'_i))| = \frac{d_k}{d_{k+1}}\frac{d_0}{d_k} = \frac{d_0}{d_{k+1}}$ for $i = 1, 2, \ldots, d_{k+1}$. Now

$$\pi_B(w_{k+1}) = \pi_B(z'_1)\pi_B(z'_2)\cdots\pi_B(z'_{d_{k+1}})$$

is the factorization of $\pi_B(w_{k+1}) = \pi_B(w'_{k+1})$ into d_{k+1} equal length $(= \frac{d_0}{d_{k+1}})$ blocks. Since $\mathrm{alph}(\pi_C(z'_i)) = \mathrm{alph}(\pi_C(u'_{\sigma(i)}))$ for all $i = 1, 2, \ldots, d_{k+1}$ and $\pi_B(w_{k+2}) = \pi_B(w'_{k+2})$, we have that

$$\pi_B(w_{k+2}) = \pi_B(u'_1)\pi_B(u'_2)\cdots\pi_B(u'_{d_{k+1}})$$

is the factorization of $\pi_B(w_{k+2})$ into d_{k+1} equal length $(= \frac{d_0}{d_{k+1}})$ blocks. Moreover, $\mathrm{alph}(\pi_B(z'_i)) = \mathrm{alph}(\pi_B(u'_{\sigma(i)}))$ for $i = 1, 2, \ldots, d_{k+1}$. Thus the condition (1) is true also for $i = k + 1$.

Surely the factorizations

$$\begin{aligned} w_{k+1} &= z'_1 z'_2 \cdots z'_{d_{k+1}}, \quad w_{k+2} = u'_1 u'_2 \cdots u'_{d_{k+1}}, \\ \pi_B(w_{k+1}) &= \pi_B(z'_1)\pi_B(z'_2)\cdots\pi_B(z'_{d_{k+1}}) \quad \text{and} \\ \pi_B(w_{k+2}) &= \pi_B(u'_1)\pi_B(u'_2)\cdots\pi_B(u'_{d_{k+1}}) \end{aligned}$$

satisfy also the condition (2). The induction is thus extended and the proof is now complete. □

Remark 4.14. Let us apply the previous lemma; choose the parameters values $d_i = n^{r-i+1}k$ for $i = 1, 2, \ldots, r$ where $k, n \in \mathbb{N}_+$. Then

$$\begin{aligned} |A| &= d_0 d_1^2 d_2^2 \cdots d_r^2 = n^{r+1}k(n^r k)^2(n^{r-1}k)^2\cdots(nk)^2 \\ &= n^{r+1}n^{2(r+(r-1)+\cdots+2+1)}k^{2r+1} \\ &= n^{(r+1)+2\frac{r(r+1)}{2}}k^{2r+1} = n^{r^2+2r+1}k^{2r+1} \\ &= n^{(r+1)^2}k^{2r+1}. \end{aligned}$$

The next theorem is of fundamental importance to our further considerations. It combines the results of Theorem 4.10 and Lemma 4.13.

Theorem 4.15. *Let α be a word and $k \geq 2$, $n \geq 1$, and $q \geq 2$ integers such that*

(1) $|\mathrm{alph}(\alpha)| \geq r_q n^{(q-1)^2 s_q} k^{(2q-3)s_q}$; *and*

(2) $|\alpha|_a \leq q$ *for each* $a \in \mathrm{alph}(\alpha)$

with r_q and s_q as in Theorem 4.10. *Then there exist $B \subseteq \mathrm{alph}(\alpha)$, $p \in \{1, 2, \ldots, q\}$ and a factorization $\alpha = \alpha_1 \alpha_2 \cdots \alpha_p$ for which*

(3) $|B| = n^{p-1}k$;

(4) $B \subseteq \mathrm{alph}(\alpha_i)$ *and the elements of B are independent with respect to $\prec_{\alpha_i}$ for $i = 1, 2, \ldots, p$; and*

(5) *for any $i \in \{1, 2, \ldots, p-1\}$, if $(\alpha_i)_B = z_1 z_2 \cdots z_{n^{p-i}k}$ is the factorization of $(\alpha_i)_B$ into $n^{p-i}k$ equal length $(= n^{i-1})$ blocks and $(\alpha_{i+1})_B = u_1 u_2 \cdots u_{n^{p-i-1}k}$ the factorization of $(\alpha_{i+1})_B$ into n^{p-i-1} equal length $(= n^i)$ blocks, then for each $j_1 \in \{1, 2, \ldots, n^{p-i}k\}$, there exists $j_2 \in \{1, 2, \ldots, n^{p-i-1}k\}$ such that* $\mathrm{alph}(z_{j_1}) \subseteq \mathrm{alph}(u_{j_2})$.

Proof. Since the conditions (1) and (2) hold, Theorem 4.10 implies that there exists $A \subseteq \mathrm{alph}(\alpha)$ with $|A| \geq n^{(q-1)^2} k^{2q-3}$ and $p \in \{1, 2, \ldots, q\}$ as well as words $\alpha_1, \alpha_2, \ldots, \alpha_p$ such that $\alpha = \alpha_1 \alpha_2 \cdots \alpha_p$ and for all $i \in \{1, 2, \ldots, p\}$, the word $(\alpha_i)_A$ is a permutation of A.

If $p = 1$, any set $B \subseteq A$ of cardinality k satisfies (3), (4) and (5). Analogously, if $p = 2$, any set $B \subseteq A$ of cardinality $n\,k$ satisfies the claims of our theorem.

Suppose that $p \geq 3$. Choose a subset A' of A such that $|A'| = n^{(p-1)^2} k^{2p-3}$. In Lemma 4.13, choose parameters as follows: $A = A'$, $r = p - 2$, $w_i = \alpha_{i+1}$ for $i = 1, 2, \ldots, p-1$, $d_0 = n^{p-1}k$, and $d_j = n^{p-1-j}k$ for $j = 1, 2, \ldots, p-2$. Then, by Lemma 4.13, there exists a subset B of A' of cardinality $|B| = d_0 = n^{p-1}k$ such that

(∗) for any $i \in \{2, 3, \ldots, p-1\}$, if $(\alpha_i)_B = x_1 x_2 \cdots x_{n^{p-i}k}$ is the factorization of $(\alpha_i)_B$ and $(\alpha_{i+1})_B = y_1 y_2 \cdots y_{n^{p-i}k}$ is the factorization of $(\alpha_{i+1})_B$ into $n^{p-i}k$ equal length $(= \frac{d_0}{d_{i-1}} = n^{i-1})$ blocks, then $\forall\, j \in \{1, 2, \ldots, n^{p-i}k\}$, there exists $j' \in \{1, 2, \ldots, n^{p-i}k\}$ such that $\mathrm{alph}(x_j) = \mathrm{alph}(y_{j'})$.

Since $B \subseteq A$, the elements of B are independent with respect to $\prec_{\alpha_i}$ for $i = 1, 2, \ldots, p$. Let $i \in \{1, 2, \ldots, p-1\}$. If $i = 1$, then certainly the claim in (5) holds, since the factorization of $\pi_B(\alpha_1)$ consists of $n^{p-1}k$ one symbol blocks.

Now suppose that $i \in \{2, 3, \ldots, p-1\}$. Let $(\alpha_i)_B = z_1 z_2 \cdots z_{n^{p-i}k}$ be the factorization of $(\alpha_i)_B$ and $(\alpha_{i+1})_B = y_1 y_2 \cdots y_{n^{p-i}k}$ of $(\alpha_{i+1})_B$ into $n^{p-i}k$ equal length blocks and $(\alpha_{i+1})_B = u_1 u_2 \cdots u_{n^{p-i-1}k}$ be the factorization of $(\alpha_{i+1})_B$ into $n^{p-i-1}k$ equal length blocks. Let $j_1 \in \{1, 2, \ldots, n^{p-i}k\}$. By the property (∗), there exists $j' \in \{1, 2, \ldots, n^{p-i}k\}$ such that $\mathrm{alph}(z_{j_1}) = \mathrm{alph}(y_{j'})$.

Since $n^{p-i-1}k$ divides $n^{p-i}k$, there exist $j_2 \in \{1, 2, \ldots, n^{p-1-i}k\}$ such that $\text{alph}(y_{j'}) \subseteq \text{alph}(u_{j_2})$. Then $\text{alph}(z_{j_1}) \subseteq \text{alph}(u_{j_2})$. The proof is now complete. □

The previous theorem clearly implies that property (P2) statet in Section 3 holds.

5 Construction and analysis of the nested multicollision attack

In this section, we shall supplement the steps of the Nested Multicollision Attack Schema so that a detailed description and analysis of a probabilistic multicollision attack procedure is possible.

5.1 The attack as a statistical experiment

Suppose that $f : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}^n$ is a compression function and $l \in \mathbb{N}_+$. Assume furthermore that we have fixed a set $Act \subseteq \mathbb{N}_l$ of so-called *active indices*. Let $\tau \in \mathbb{N}_l^+$ be a word such that $\text{alph}(\tau)$ contains exactly one element, say t, which is an active index. Finally, let $\omega \in \{0,1\}^m$ be a given constant message block.

A *basic birthday attack on f_τ with active index t and initial value h*, denoted by $BBA(f_\tau, t, h)$ is understood to be a statistical (probabilistic) experiment carried out as follows.

(1) Generate a set $R \subseteq \{0,1\}^m$ of $2^{\frac{n}{2}}$ random message blocks.

(2) Let

$$S = \{u_1 u_2 \cdots u_l \mid u_t \in R \text{ and } \forall i \in \mathbb{N}_l \setminus \{t\} : u_i = \omega\}.$$

(3) For each $u \in S$, compute the value $f_\tau(h, u)$ to find message blocks $x, y \in R$, $x \neq y$, and the respective collision value h' such that

$$f_\tau(h, \omega^{t-1} x \omega^{l-t}) = f_\tau(h, \omega^{t-1} y \omega^{l-t}) = h'.$$

The probability $\tilde{p}$ that $BBA(f_\tau, t, h)$ yields a collision is approximately equal to 0.4 (for details, see for instance [14, 16]). In an (*extended*) *birthday attack* on f_τ with active index t and initial value h, (abbreviated $EBA(f_\tau, t, h)$) one or more basic birthday attacks are carried out one after another until a collision is found. Thus in an extended birthday attack a collision is always found with probability equal to one. The expected number $\tilde{a}$ of BBAs in an EBA is obviously equal to

$1/\tilde{p}$. As mentioned above, $\tilde{p} \approx 0.4$, so we have $\tilde{a} \approx 2.5$. Thus the expected number of queries on f in $EBA(f_\tau, t, h)$ is equal to $\tilde{a}|\tau|2^{\frac{n}{2}}$.

Let now α be a word over the alphabet $\mathbb{N}_l$ and Act be the set of $r \in \mathbb{N}_+$ active indices $a_1, a_2, \ldots, a_r$ such that $a_1 \prec_\alpha a_2 \prec_\alpha \cdots \prec_\alpha a_r$. Suppose furthermore that $\alpha = \alpha_1\alpha_2\cdots\alpha_r$ is a factorization of α such that for each $i \in \{1, 2, \ldots, r\}$, all occurrences of the symbol a_i in α lie in α_i. In our construction (see Lemma 5.1), a sequence

$$EBA(f_{\alpha_1}, h_0, a_1), EBA(f_{\alpha_2}, h_1, a_2), \ldots, EBA(f_{\alpha_r}, h_{r-1}, a_r)$$

of extended birthday attacks is executed. Above h_0 is the initial value and for each $i \in \{1, 2, \ldots, r\}$, during the execution of $EBA(f_{\alpha_i}, h_{i-1}, a_i)$, values $h_i \in \{0,1\}^n$ and distinct message blocks $x_{a_i}, y_{a_i} \in \{0,1\}^m$ are found such that

$$h_i = f_{\alpha_i}(h_{i-1}, \omega^{a_i-1}x_{a_i}\omega^{l-a_i}) = f_{\alpha_i}(h_{i-1}, \omega^{a_i-1}y_{a_i}\omega^{l-a_i}).$$

The collision value h_i of $EBA(f_{\alpha_i}, h_{i-1}, a_i)$ serves as the initial value to the attack $EBA(f_{\alpha_{i+1}}, h_i, a_{i+1})$ for $i = 1, 2, \ldots, r-1$. We may assume that the EBA's above are statistically independent, so the expected number of BBA's in the sequence is $\tilde{a} \cdot r$. We may also deduce that the expected number of queries on the total sequence is equal to $\tilde{a}|\alpha_1\alpha_2\cdots\alpha_r|2^{\frac{n}{2}}$. Obviously the set

$$M = \{u_1u_2\cdots u_l \mid \forall i \in \{1, 2, \ldots, r\} : u_{a_i} \in \{x_{a_i}, y_{a_i}\} \\ \wedge\ \forall i \in \mathbb{N}_l \setminus \mathrm{Act} : u_i = \omega\}$$

is a 2^r-collision in f_α with initial value h_0. If we above choose $\alpha_i = a_i$ for $i = 1, 2, \ldots, r$, we can interpret Joux's 2^r-collision attack to be a special case of our construction: certainly the complexity of this attack is $\tilde{a}\, r\, 2^{\frac{n}{2}}$.

The time is now ripe to augment the first three steps in the schema *NMCAS*. Call the expanded plan of action *Nested Multicollision Attack* (*NMCA*).

Procedure *NMCA*

Input: A q-bounded ($q \in \mathbb{N}$, $q \geq 2$) generalized iterated hash function $H_{\hat{\alpha},f}$, initial value $h_0 \in \{0,1\}^n$, integer $r \in \mathbb{N}_+$.
Output: A 2^r-collision in $H_{\hat{\alpha},f}$.
Step 1: Let $l = r_q n^{(q-1)^2 s_q} r^{(2q-3)s_q}$ where r_q and s_q are parameters defined in Theorem 4.10. Let $\alpha = \alpha_l$ where α_l is the lth element of the sequence $\hat{\alpha}$. Write α in the form $\alpha = i_1 i_2 \cdots i_s$, where $s \in \mathbb{N}_+$ and $i_j \in \mathbb{N}_l$ for $j = 1, 2, \ldots, s$.
Step 2: Let $Act = B$, $|B| = n^{p-1}r$, be the set of active indices, where $B \subseteq \mathbb{N}_l = \{1, 2, \ldots, l\}$ and $p \in \{1, 2, \ldots, q\}$ are as in Theorem 4.15, when the parameter $k = r$.

Step 3: Let $\alpha = \beta_1\beta_2\cdots\beta_p$ the factorization of α such that the words $\beta_1, \beta_2, \ldots,$ β_p have the same properties as the words $\alpha_1, \alpha_2, \ldots, \alpha_p$, respectively, in Theorem 4.15, when $k = r$.

Note that, in Step 2 above, no algorithm to find the set B is specified. In the trivial case, given $l = r_q n^{(q-1)^2 s_q} r^{(2q-3)s_q}$, one could check all the $\binom{l}{n^{p-1}r}$ subsets of size $n^{p-1}r$ of $\{1, 2, \ldots, l\}$. This certainly can be carried out in polynomial time with respect to n.

5.2 The two phases of the attack

Our next task is to show that Step 4 in *NMCAS* is feasible, i.e., the multicollision can be constructed so that the expected number of queries on f is not too large. The next lemma is an extended version of Theorem 5.1 in [16].

Lemma 5.1. *Let α be a word over the alphabet $\mathbb{N}_l$, $r \in \mathbb{N}_+$, and $a_1, a_2, \ldots, a_r$ in* alph(α)*, symbols such that $a_1 \prec_\alpha a_2 \prec_\alpha \ldots \prec_\alpha a_r$. Let furthermore $\alpha = \alpha_1\alpha_2\cdots\alpha_r$ be a factorization of α such that for each $i \in \{1, 2, \ldots, r\}$, all occurrences of the symbol a_i in α lie in α_i. Given an initial value $h_0 \in \{0,1\}^n$, we can, with probability equal to one, find message block sets $M_1, M_2, \ldots, M_l \subseteq \{0,1\}^m$ as well as values $h_1, h_2, \ldots, h_r \in \{0,1\}^n$ such that*

(1) *$M_b = \{\omega\}$ for each $b \in \mathbb{N}_l \setminus A$, where $A = \{a_1, a_2, \ldots, a_r\}$;*

(2) *$M_{a_i} = \{u_i, u_i'\}$, where $u_i \neq u_i'$ for each $i \in \{1, 2, \ldots, r\}$;*

(3) *for each $i \in \{1, 2, \ldots, r\}$ the set $M = M_1 \cdot M_2 \cdots M_l$ is a 2-collision in f_{α_i} with initial value h_{i-1} and a 2^i-collision in $f_{\alpha_1\alpha_2\cdots\alpha_i}$ such that $\forall u, u' \in M$*

$$h_i = f_{\alpha_i}(h_{i-1}, u) = f_{\alpha_i}(h_{i-1}, u') \quad \text{and}$$
$$f_{\alpha_1\alpha_2\cdots\alpha_i}(h_0, u) = f_{\alpha_1\alpha_2\cdots\alpha_i}(h_0, u').$$

Moreover, the expected number of queries on f needed to carry out the task is $\tilde{a}|\alpha|2^{\frac{n}{2}}$.

Proof. Let initially $M_i = \{\omega\}$ for $i = 1, 2, \ldots, l$ and $M = M_1 \cdot M_2 \cdots M_l$. Proceed by induction on r. Suppose that, given the initial value $h_0 \in \{0,1\}^n$ we are, with probability equal to one, able to find message block sets $M_{a_i} = \{u_i, u_i'\}, u_i \neq u_i', i = 1, 2, \ldots, r-1$, as well as values $h_1, h_2, \ldots, h_{r-1} \in \{0,1\}^n$ such that after updating $M := M_1 \cdot M_2 \cdots M_l$ the following holds: for each

$i \in \{1, 2, \ldots, r-1\}$ the set M is a 2-collision in f_{α_j} with initial value h_{j-1} such that

$$\forall u, u' \in M : h_j = f_{\alpha_j}(h_{j-1}, u) = f_{\alpha_j}(h_{j-1}, u').$$

Furthermore, assume that the expected number of queries on f is equal to $\tilde{a}|\alpha_1\alpha_2 \cdots \alpha_{r-1}| \cdot 2^{\frac{n}{2}}$. Replace in $M = M_1 \cdot M_2 \cdots M_l$ the message block set $M_{a_r} = \{\omega\}$ with a set T_{a_r} of $2^{\frac{n}{2}}$ random message blocks and denote the attained set by T. Among the messages, a 2-collision in f_{α_i} with initial value h_{i-1} is searched for. Reasoning exactly as in the beginning of the previous Subsection 5.1, we deduce that, to find a collision in f_{α_r}, the expected number of times that the generation of the set T_{a_r} of $2^{\frac{n}{2}}$ random message blocks has to be repeated is $\tilde{a}$. Thus, to find a collision in f_{α_r}, the expected number of queries on f is $\tilde{a}|\alpha_i|2^{\frac{n}{2}}$. Note two things in the construction of the collision in f_{α_i}:

(i) only message blocks from those sets M_j for which $j \in \text{alph}(\alpha_r)$ are used; and

(ii) for each $a \in \text{alph}(\alpha_{a_r})$, if $a \neq a_r$, then $M_a = \{\omega\}$.

Let $x, y \in T$, $x \neq y$, be such that $f_{\alpha_r}(h_{r-1}, x) = f_{\alpha_r}(h_{r-1}, y)$. Let $x = x_1x_2 \cdots x_l$ and $y = y_1y_2 \cdots y_l$, where $x_i, y_i \in \{0,1\}^m$, for all $i = 1, 2, \ldots, l$. By the properties (i) and (ii) above, $x_{a_r} \neq y_{a_r}$. Choose $u_r = x_{a_r}$ and $u'_r = y_{a_r}$. Let $M_{a_r} = \{u_i, u'_i\}$ and $h_r = f_{\alpha_r}(h_{r-1}, x)$. Update $M = M_1 \cdot M_2 \cdots M_l$ and deduce that $\forall u, u' \in M$

$$\begin{aligned} h_r &= f_{\alpha_r}(h_{r-1}, u) = f_{\alpha_r}(h_{r-1}, u') \\ &= f_{\alpha_1\alpha_2\cdots\alpha_r}(h_0, u) = f_{\alpha_1\alpha_2\cdots\alpha_r}(h_0, u'). \end{aligned}$$

Obviously, M is a 2-collision in f_{α_r} with initial value h_{r-1} and a 2^r-collision in $f_{\alpha_1\alpha_2\cdots\alpha_r}$. The expected number of queries on f is $\tilde{a}|\alpha_1\alpha_2 \cdots \alpha_r|2^{\frac{n}{2}}$ in all. The induction is now extended. □

We can now top up the fourth step of *NMCAS*.

Step 4 of *NMCA*: Let $M_1, M_2, \ldots, M_l$ be as in Lemma 5.1.

Our next result implies that in Step 5 of *NMCAS* for any $r \in \{2, 3, \ldots, p\}$, the set $\mathcal{C}_i$ can be constructed from $\mathcal{C}_{i-1}$ feasibly, i.e., so that the expected number of queries on f is again not too high. Recall the definition of $\bar{u}$: if $\alpha = a_1a_2 \cdots a_s$ and $u = u_1u_2 \cdots u_l$ are words such that $a_i \in \mathbb{N}_l$ for $i = 1, 2, \ldots, s$ and $u_j \in \{0,1\}^m$ for $j = 1, 2, \ldots, l$, then $\bar{u}(\alpha) = u_{a_1}u_{a_2} \cdots u_{a_s}$.

Lemma 5.2. *Let α be a word over the alphabet $\mathbb{N}_l$, d and r positive integers, $A \subseteq \text{alph}(\alpha)$ a set of cardinality $|A| = dnr$, and $\alpha = \beta_1\beta_2\cdots\beta_{nr}\gamma_1\gamma_2\cdots\gamma_r$ a factorization of α with the following properties.*

(1) $A \subseteq \text{alph}(\beta) \cap \text{alph}(\gamma)$ *where* $\beta = \beta_1\beta_2\cdots\beta_{nr}$ *and* $\gamma = \gamma_1\gamma_2\cdots\gamma_r$;

(2) $|\text{alph}(\beta_i) \cap A| = d$ *for* $i = 1, 2, \ldots, nr$, *and* $|\text{alph}(\gamma_j) \cap A| = nd$ *for* $j = 1, 2, \ldots, r$; *and*

(3) *for each* $i \in \{1, 2, \ldots, nr\}$ *there exists* $j \in \{1, 2, \ldots, r\}$ *such that* $\text{alph}(\beta_i) \cap A \subseteq \text{alph}(\gamma_j) \cap A$.

Moreover, let $u_1, u_1', u_2, u_2', \ldots, u_{nr}, u_{nr}' \in \{0,1\}^{ml}$ *be messages and* $h_0, h_1, \ldots, h_{nr} \in \{0,1\}^n$ *be values such that for each* $i \in \{1, 2, \ldots, nr\}$:

(4) $\forall b \in \mathbb{N}_l \setminus A : \bar{u}_i(b) = \bar{u}_i'(b) = \omega$; *and*

(5) $\bar{u}_i(\beta_i) \neq \bar{u}_i'(\beta_i)$ *and* $h_i = f_{\beta_i}(h_{i-1}, u_i) = f_{\beta_i}(h_{i-1}, u_i')$.

Then the set S of all messages $u \in \{0,1\}^{ml}$ such that for each $b \in \mathbb{N}_l \setminus A$: $\bar{u}(b) = \omega$ and for each $i \in \{1, 2, \ldots, nr\}$: $\bar{u}(\beta_i) \in \{\bar{u}_i(\beta_i), \bar{u}_i'(\beta_i)\}$ is well-defined and satisfies for each $i \in \{1, 2, \ldots, nr\}$ and $u \in S$ the equality $h_i = f_{\beta_i}(h_{i-1}, u)$. Moreover we can, with probability equal to one, find messages $v_1, v_1', v_2, v_2', \ldots, v_r, v_r'$ in S and values $h_0', h_1', \ldots h_r'$, $h_0' = h_{nr}$, such that for each $j \in \{1, 2, \ldots, r\}$:

(6) $\bar{v}_j(\gamma_j) \neq \bar{v}_j'(\gamma_j)$ *and* $h_j' = f_{\gamma_j}(h_{j-1}', v_j) = f_{\gamma_j}(h_{j-1}', v_j')$.

The expected number of queries on f needed to carry out the task is $\tilde{a}|\gamma|2^{\frac{n}{2}}$. Finally, the set T of all messages $v \in \{0,1\}^{ml}$ such that for each $b \in \mathbb{N}_l \setminus A$: $\bar{v}(b) = \omega$ and for each $j \in \{1, 2, \ldots, r\}$: $\bar{v}(\gamma_j) \in \{\bar{v}_j(\gamma_j), \bar{v}_j'(\gamma_j)\}$ is a well-defined subset of S and forms a nontrivial 2^r-collision on f_α with initial value h_0.

Proof. Note first that since $|A| = dnr$, $A \subseteq \text{alph}(\beta)$, and $|\text{alph}(\beta_i) \cap A| = d$ for each $i \in \{1, 2, \ldots, nr\}$, the indexed family of sets $\{\text{alph}(\beta_i) \cap A\}_{i=1}^{nr}$ forms a partition of A. With analogous reasoning, $\{\text{alph}(\gamma_j) \cap A\}_{j=1}^{r}$ is a partition of A, too.

Let now $x_i \in \{u_i, u_i'\}$ for $i = 1, 2, \ldots, nr$. Consider the sequence $\bar{x}_1(\beta_1)$, $\bar{x}_2(\beta_2)$, $\ldots$, $\bar{x}_{nr}(\beta_{nr})$. Define $t_1, t_2, \ldots, t_l \in \{0,1\}^m$ as follows. For each $b \in \mathbb{N}_l \setminus A$, let $t_b = \omega$. For each $a \in A$ and $i \in \{1, 2, \ldots, nr\}$, if $a \in \text{alph}(\beta_i) \cap A$, then $t_a = \bar{x}_i(a)$. Since $\{\text{alph}(\beta_i) \cap A\}_{i=1}^{nr}$ is a partition of A, the message block t_a is uniquely determined. Thus the sequence $\bar{x}_1(\beta_1), \bar{x}_2(\beta_2), \ldots, \bar{x}_{nr}(\beta_{nr})$ uniquely defines the message $t_1 t_2 \cdots t_l$. We deduce that the set S is well-defined.

Consider now the sets $\{\bar{u}(\gamma_1)|u \in S\}$, $\{\bar{u}(\gamma_2)|u \in S\}, \ldots, \{\bar{u}(\gamma_r) \,|\, u \in S\}$. Since $\{\mathrm{alph}(\gamma_j) \cap A\}_{j=1}^{r}$ is a partition of A and the property (3) holds, the cardinality of the set $\{\bar{u}(\gamma_j) \,|\, u \in S\}$ is 2^n for each $j \in \{1, 2, \ldots, r\}$. Furthermore, since $\gamma = \gamma_1\gamma_2\cdots\gamma_r$, the equality

$$\{\bar{u}(\gamma) \,|\, u \in S\} = \{\bar{u}(\gamma_1) \,|\, u \in S\}\{\bar{u}(\gamma_2) \,|\, u \in S\}\cdots\{\bar{u}(\gamma_r) \,|\, u \in S\}$$

holds, so the cardinality of the set $\{\bar{u}(\gamma) \,|\, u \in S\}$ is 2^{nr}.

Let $u \in S$. Then

$$f_\beta(h_0, u) = f^+(h_0, \bar{u}(\beta)) = f^+(h_0, \bar{u}(\beta_1)\bar{u}(\beta_2)\cdots\bar{u}(\beta_{nr})) = h_{nr}\,.$$

Thus S is a 2^{nr}-collision in f_β with initial value h_0.

By assumption, $h'_0 = h_{nr}$. Continue by induction; assume that k is in $\{1, 2, \ldots, r-1\}$ and with probability equal to one, messages $v_1, v'_1, v_2, v'_2, \ldots, v_k, v'_k$ in S and values $h'_1, h'_2, \ldots, h'_k$ in $\{0,1\}^n$ have been found such that for each $j \in \{1, 2, \ldots, k\}$

$$\bar{v}_j(\gamma_j) \neq \bar{v}'_j(\gamma_j) \quad \text{and} \quad h'_j = f_{\gamma_j}(h'_{j-1}, v_j) = f_{\gamma_j}(h'_{j-1}, v'_j).$$

Furthermore, the expected number of queries on f is $\tilde{a}|\gamma_1\gamma_2\cdots\gamma_k|2^{\frac{n}{2}}$. Since, for each $u \in S$, the equality

$$f_{\gamma_{k+1}}(h'_k, u) = f^+(h'_k, \bar{u}(\gamma_{k+1}))$$

holds and the cardinality of the set $\{\bar{u}(\gamma_{k+1}) \,|\, u \in S\}$ is 2^n. Thus we can, choosing randomly from the set S message sets of cardinality $2^{\frac{n}{2}}$ and reasoning exactly as in the proof of Lemma 5.1, with probability equal to one, find messages v_{k+1}, v'_{k+1} in $\{0,1\}^{ml}$ and a value h'_{k+1} in $\{0,1\}^n$ such that $\bar{v}_{k+1}(\gamma_{k+1}) \neq \bar{v}'_{k+1}(\gamma_{k+1})$ and $h'_{k+1} = f_{\gamma_{k+1}}(h'_k, v_{k+1}) = f_{\gamma_{k+1}}(h'_k, v'_{k+1})$. The expected number of queries on f is certainly $\tilde{a}|\gamma_{k+1}|2^{\frac{n}{2}}$.

The induction is now extended and messages $v_1, v'_1, v_2, v'_2, \ldots, v_r, v'_r$ in S and values $h'_0, h'_1, \ldots, h'_r$ in $\{0,1\}^n$ satisfying (6) found with expected number $\tilde{a}|\gamma|2^{\frac{n}{2}}$ of queries on f. The task is successful with probability one.

Reasoning as with the set S and noting that $v_j, v'_j \in S$ for all $j \in \{1, 2, \ldots, r\}$, it is straightforward to see that T is a well-defined subset of S. Since $\bar{v}_j(\gamma_j) \neq \bar{v}'_j(\gamma_j)$ for each $j \in \{1, 2, \ldots, r\}$ and $\bar{v}_j(b) = \bar{v}'_j(b)$ for all $b \in \mathbb{N}_l \setminus A$, the cardinality of T is 2^r. Certainly $f_\alpha(h_0, u) = h'_r$ for each $u \in T$. The proof is now complete. □

The following theorem combines the results of the two previous lemmata; we verify that Step 5 in *NMCAS* can be carried out in a feasible fashion without consuming too much resources.

Theorem 5.3. *Let α be a word over the alphabet $\mathbb{N}_l$, r and p positive integers, A a subset of the alphabet* $\mathrm{alph}(\alpha)$ *of cardinality $|A| = n^{p-1}r$, and $\alpha = \alpha_1\alpha_2\cdots\alpha_p$ a factorization of α such that for each $i \in \{1, 2, \ldots, p\}$, the elements of A form a chain in the partially ordered set* $(\mathrm{alph}(\alpha), \prec_{\alpha_i})$ *(i.e., the elements of A are independent with respect to $\prec_{\alpha_i}$). Assume furthermore that for each $i \in \{1, 2, \ldots, p\}$, there exists a factorization $\alpha_i = \alpha_{i1}\alpha_{i2}\cdots\alpha_{i,n^{p-i}r}$ of the word α_i such that the following conditions are satisfied.*

(1) $|\mathrm{alph}(\alpha_{ij}) \cap A| = n^{i-1}$ *for each $i \in \{1, 2, \ldots, p\}$ and $j \in \{1, 2, \ldots, n^{p-i}r\}$; and*

(2) *for all $i \in \{1, 2, \ldots, p\}$ and $j \in \{1, 2, \ldots, n^{p-i}r\}$ there exists $k \in \{1, 2, \ldots, n^{p-i-1}r\}$ such that* $\mathrm{alph}(\alpha_{ij}) \cap A$ *is a subset of* $\mathrm{alph}(\alpha_{i+1,k}) \cap A$.

Then, given an initial value $h_0 \in \{0,1\}^n$ we can, with probability equal to one, find a nontrivial 2^r-collision in f_α. Moreover, the expected number of queries on f_α needed to carry out the task is $\tilde{a}|\alpha|2^{\frac{n}{2}}$.

Proof. We first apply Lemma 5.1 to generate a $2^{n^{p-1}r}$-collision set B_1 on f_{α_1} and then, by using Lemma 5.2 repeatedly, show that there exists a $2^{n^{p-i}r}$-collision B_i on $f_{\alpha_1\alpha_2\cdots\alpha_i}$ for $i = 2, 3, \ldots, p$ such that $B_1 \supsetneq B_2 \supsetneq \cdots \supsetneq B_p$.

In Lemma 5.1, choose the parameters as follows: α is equal to α_1 and r is equal to $n^{p-1}r$. Let $A = \{a_1, a_2, \ldots, a_{n^{p-1}r}\}$ and $a_1 \prec_\alpha a_2 \prec_\alpha \ldots \prec_\alpha a_{n^{p-1}r}$ $i = 1, 2, \ldots, p$. Certainly these assumptions can be made.

Then $\alpha_1 = \alpha_{11}\alpha_{12}\cdots\alpha_{1,n^{p-1}r}$ is a factorization of α_1 such that all occurrences of the symbol a_j in α_1 lie in α_{1j}, for each $j \in \{1, 2, \ldots, n^{p-1}r\}$. Let $h_0 \in \{0,1\}^n$ be given. Applying Lemma 5.1, one can, with probability equal to one and with expected number $\tilde{a}|\alpha_1|2^{\frac{n}{2}}$ of queries on f, find message block sets $M_1, M_2, \ldots, M_l \subseteq \{0,1\}^m$ as well as values $h_1, h_2, \ldots, h_{n^{p-1}r}$ such that

(a) $M_b = \{\omega\}$ for each $b \in \mathbb{N}_l \setminus A$;

(b) $M_{a_i} = \{w_i, w_i'\}$, where $w_i \neq w_i'$ for each $i \in \{1, 2, \ldots, n^{p-1}r\}$; and

(c) for each $i \in \{1, 2, \ldots, n^{p-1}r\}$, the set $M = M_1M_2\cdots M_l$ is a 2-collision in $f_{\alpha_{1i}}$ and a 2^i-collision in $f_{\alpha_1\alpha_2\cdots\alpha_i}$ with initial value h_{i-1} such that for each u and u' in M:

$$h_i = f_{\alpha_{1i}}(h_{i-1}, u) = f_{\alpha_{1i}}(h_{i-1}, u') \quad \text{and}$$
$$f_{\alpha_{11}\alpha_{12}\cdots\alpha_{1i}}(h_0, u) = f_{\alpha_{11}\alpha_{12}\cdots\alpha_{1i}}(h_0, u').$$

Obviously the set $B_1 = M$ is a $2^{n^{p-1}r}$-collision in f_{α_1} with initial value h_0. The creation of B_1 was carried out by a statistical process which succeeds with

probability equal to one; in the process the expected number of queries on f is $\tilde{a}|\alpha_1|2^{\frac{n}{2}}$.

Choose the parameters of Lemma 5.2 as follows. Let β be α_1, γ be α_2 and r be $n^{p-1}r$. Let d be equal to 1, β_i equal to α_{1i} for $i = 1, 2, \ldots, n^{p-1}r$, and γ_j equal to α_{2j} for $j = 1, 2, \ldots, n^{p-2}r$. Then the assumptions of Theorem 5.3 for α_1 and α_2 imply that all the assumptions of Lemma 5.2 (with parameters chosen as above) are valid. Thus we can, with a probability equal to one and expected number $\tilde{a}|\alpha_2|2^{\frac{n}{2}}$ of queries on f, find messages $v_1, v'_1, v_2, v'_2, \ldots, v_{n^{p-2}r}, v'_{n^{p-2}r}$ in $\{0,1\}^{ml}$ and values $h'_0, h'_1, \ldots, h'_{n^{p-2}r}$ in $\{0,1\}^n$, $h'_0 = h_{n^{p-1}r}$, such that for each $j \in \{1, 2, \ldots, n^{p-2}r\}$, $\forall b \in \mathbb{N}_l \setminus A : \bar{v}_j(b) = \bar{v}'_j(b) = \omega$ and $\bar{v}_j(\alpha_{2j}) \neq \bar{v}'_j(\alpha_{2j})$ and $h'_j = f_{\alpha_{2j}}(h'_{j-1}, v_j) = f_{\alpha_{2j}}(h'_{j-1}, v'_j)$. The set S of Lemma 5.2 is clearly our set $B_1 = M$. Choose B_2 to be the set T guaranteed by Lemma 5.2. Then $B_2 \subsetneq B_1$ is a (nontrivial) $2^{n^{p-2}r}$-collision in $f_{\alpha_1\alpha_2}$ with initial value h_0.

Continue by induction and let $k \in \{2, 3, \ldots, p-1\}$. Let the words $x_1, x'_1, x_2, x'_2, \ldots, x_{n^{p-k}r}, x'_{n^{p-k}r} \in \{0,1\}^{ml}$ and values $d_0, d_1, \ldots, d_{n^{p-k}r} \in \{0,1\}^n$ be such that for each $i \in \{1, 2, \ldots, n^{p-k}r\}$, $\forall b \in \mathbb{N}_l \setminus A : \bar{x}_i(b) = \bar{x}'_i(b) = \omega$ and $\bar{x}_i(\alpha_{ki}) \neq \bar{x}'_i(\alpha_{ki})$ and $d_i = f_{\alpha_{ki}}(d_{i-1}, x_i) = f_{\alpha_{ki}}(d_{i-1}, x'_i)$. Let B_k be the set of all messages $u \in \{0,1\}^{ml}$ such that for each $b \in \mathbb{N}_l \setminus A$: $\bar{u}(b) = \omega$ and for each $j \in \{1, 2, \ldots, n^{p-k}r\}$: $\bar{u}(\alpha_{kj})$ is in $\{\bar{x}(\alpha_{kj}), \bar{x}'(\alpha_{kj})\}$. Suppose that B_k is a subset of B_{k-1} and that B_k is a $2^{n^{p-k}r}$-collision in $f_{\alpha_1\alpha_2\cdots\alpha_k}$ with initial value h_0. Choose the parameters of Lemma 5.2 as follows. Let d be equal to $n^{p-k-1}r$, β be equal to α_k, β_i be equal to α_{ki} for $i = 1, 2, \ldots, n^{p-k}r$, and γ_j be equal to $\alpha_{k+1,j}$ for $j = 1, 2, \ldots, n^{p-k-1}r$. By the assumptions of Theorem 5.3, all the assumptions of Lemma 5.2 are valid (with the chosen parameter values). Lemma 5.2 implies that one may, with a probability equal to one and expected number $\tilde{a}|\alpha_{k+1}|2^{\frac{n}{2}}$ of queries on f, find messages $y_1, y'_1, y_2, y'_2, \ldots, y_{n^{p-k-1}r}, y'_{n^{p-k-1}r}$ in $\{0,1\}^{ml}$ and values $d'_0, d'_1, \ldots, d'_{n^{p-k-1}r}$ in $\{0,1\}^n$, $d'_0 = d_{n^{p-k}r}$, such that for each $j \in \{1, 2, \ldots, n^{p-k-1}r\}$, $\forall b \in \mathbb{N}_l \setminus A : \bar{y}_j(b) = \bar{y}'_j(b) = \omega$ and $\bar{y}_j(\alpha_{k+1,j}) \neq \bar{y}'_j(\alpha_{k+1,j})$ and $d'_j = f_{\alpha_{k+1,j}}(d'_{j-1}, y_j) = f_{\alpha_{k+1,j}}(d'_{j-1}, y'_j)$. The set T of all messages y in $\{0,1\}^{ml}$ such that for each $b \in \mathbb{N}_l \setminus A$: $\bar{y}(b) = \omega$ and for each $j \in \{1, 2, \ldots, n^{p-k-1}r\}$: $\bar{y}(\alpha_{k+1,j})$ is in $\{\bar{y}(\alpha_{k+1,j}), \bar{y}'(\alpha_{k+1,j})\}$ is then a well-defined subset of B_k and forms a (nontrivial) $2^{n^{p-k-1}r}$-collision in $f_{\alpha_k\alpha_{k+1}}$ with initial value d_0. By the induction assumption, T is a $2^{n^{p-k-1}r}$-collision in $f_{\alpha_1\alpha_2\cdots\alpha_{k+1}}$ with initial value h_0. Choose $B_{k+1} = T$ and the induction is extended. We deduce that we can, with probability equal to one, find a nontrivial 2^r-collision in f_α with initial value h_0. The expected number of queries on f is altogether $\tilde{a}|\alpha|2^{\frac{n}{2}}$. □

The fifth step of step of *NMCAS* can be completed:

Step 5 of *NMCA*: Let $B_1, B_2, \ldots, B_p$ be as in the proof of Theorem 5.3. The theorem guarantees that, to create a 2^r-collision in f_α, the expected number of queries on f is $\tilde{a}|\alpha|2^{\frac{n}{2}}$.

Let us recapitulate our results.

Theorem 5.4. *Let m, n and q be positive integers such that $m > n$ and $q \geq 2$, f a compression function of block size m and length n, and $\hat{\alpha} = (\alpha_1, \alpha_2, \ldots)$ a q-bounded sequence of words such that* $\mathrm{alph}(\alpha_i) = \mathbb{N}_l$ *for each $i \in \mathbb{N}_+$. Then, for each $r \in \mathbb{N}_+$, there exists a 2^r-collision attack on the generalized iterated hash function $H_{\hat{\alpha},f}$ of complexity $\tilde{a}\, r_q\, n^{(q-1)^2 s_q} r^{(2q-3)s_q} 2^{\frac{n}{2}}$, where the parameters r_q and s_q are defined recursively by $r_1 = s_1 = 1$, $r_{i+1} = i^{s_i} r_i^{s_i+1}$ and $s_{i+1} = s_i^2 + 1$ for $i \in \mathbb{N}_+$.*

We wish to recall that in [9] an informal proof of the previous theorem with a different complexity and parameter definitions was given.

5.3 The case $q = 2$ and some complexity considerations

Now suppose that in the input of the procedure *NMCA* the generalized iterated hash function $H_{\hat{\alpha},f}$ is such that the sequence $\hat{\alpha} = (\alpha_1, \alpha_2, \ldots)$ is 2-bounded. Then, by Theorem 4.10, the equalities $r_2 = 2^{s_1} r_1^{s_1+1} = 2$ and $s_2 = s_1^2 + 1 = 2$ hold. By Theorem 5.4, when creating a 2^r-collision ($r \in \mathbb{N}_+$) on $H_{\hat{\alpha},f}$, the expected number of queries on f is $\tilde{a}\, r_2\, n^{(2-1)^2 s_2} r^{(2\cdot 2-3)s_2} 2^{\frac{n}{2}} = 2\,\tilde{a}\, n^2\, r^2 2^{\frac{n}{2}}$. In [16] with rigorous considerations a somewhat smaller average complexity $O(r^2 \cdot (\ln r) \cdot (n + \ln(\ln 2r)) \cdot 2^{n/2})$ was attained.

The method of Nandi and Stinson [16] guarantees that a 2^r-collision is reached regardless of the number of permutations (one or two). The method applied in this paper yields either a 2^{nr}-collision (for one permutation) or a 2^r-collision (in the case of two permutations). This leads to a somewhat rougher estimate and thus to greater complexity in this special case. It would be interesting to see if some of the techniques in [16] could also be used in the general case to lower the complexity.

Note that in *NMCA* it is possible to take also q as an input parameter. Then the procedure is of course extremely inefficient: due to the recurrence relations $r_1 = s_1 = 1$, and $r_{q+1} = q^{s_q} r_q^{s_q+1}$, $s_{q+1} = s_q^2 + 1$ for $q \in \mathbb{N}_+$, we have a procedure that is at least triple exponential with respect to q. A natural question arises whether or not in Theorem 4.10, the length of the word α could be chosen to be considerably smaller. Our opinion is that then a significantly different proof technique is needed. If Lemma 4.8 and Dilworth's Theorem (and thus the relation

between independent elements and incomparable elements in alph(α) is applied, it is difficult to imagine that a 2^r-collision in $H_{\hat{\alpha},f}$ could be constructed so that the expected number of queries on f is less than exponential with respect to q.

To implement a generalized iterated hash function, a relatively strong computing device (in automata theoretic sense) is needed. In fact, a two-way deterministic pushdown transducer seems to be an indispensable tool regardless of the way the respective compression function is realized as a computer program. This raises the question of efficiency because a two-way deterministic pushdown transducer is a much more complicated machine (and thus much more resource consuming to implement and use) than a finite state transducer which is needed to realize a traditional iterated hash function.

If a generalized iterated hash function is used, the sender has to construct the whole message before he can start to hash it. Similarly, the receiver has to have the complete message available before the sent hash value can be verified to be correct. This greatly impedes their applicability in applications where streaming data is used. Suppose that we wish to somehow avoid this restriction and start hashing the message before it has completely been formed or received. Then the message blocks occurring at the end of the message are not available when we start hashing. This causes extra restrictions on the sequence $\hat{\alpha}$ and possibly implies that multicollisions are easily found. Especially chains could be forced to form as the earlier message blocks will be used up before the message blocks in the end can be applied. As can be seen from our method, this enables a relatively fast and straightforward multicollision attack.

We also need an efficient encoding for the sequence $\hat{\alpha}$. If $\hat{\alpha}$ is complicated, which means that the hash function $H_{\hat{\alpha},f}$ is secure, then picking an element α_l from $\hat{\alpha}$ may be resource consuming. On the other hand, if $\hat{\alpha}$ is very simple and picking an element from the sequence can be done with ease, then there might be very efficient multicollision attacks against these types of hash functions.

6 Conclusion

In this paper, we have demonstrated how the analysis of multicollisions in iterated hash functions can be done with the use of word combinatorics. We have also given some new results and settled some inaccuracies in the previous results concerning multicollisions in generalized iterated hash functions. We have also brought these results into a unified and well established theoretical framework, which should make further investigation of the theory of iterated hash functions easier.

The next step in the research could be to investigate the possibilities of generating words, which have desirable properties in the context of multicollisions.

We could also categorize words and whole languages with respect to their performance in the iterative structure. Also the even more general types of iterated hash functions presented in [16] and [9] could be brought into this framework and further analyzed.

Acknowledgments. The authors wish to thank the anonymous referees for carefully reading the manuscript, for expert comments that improved the text considerably, and patience towards the immature first version of the paper. Furthermore, we wish to thank the participants of the Oulu University Hash Function Seminar, who provided valuable and insightful comments on our work.

Bibliography

[1] Alfred V. Aho and Neil J. A. Sloane, *Some doubly exponential sequences*, Fibonacci Quarterly 11 (1973), 429–437.

[2] Sabra S. Anderson, *Graph Theory and Finite Combinatorics*. Markham, Chicago, 1970.

[3] Jean-Sébastien Coron, Yevgeniy Dodis, Cécile Malinaud and Prashant Puniya, *Merkle–Damgård revisited: How to construct a hash function*. Advances in Cryptology – CRYPTO '05 (Victor Shoup, ed.), Lecture Notes in Computer Science 3621, pp. 430–448. Springer, 2005.

[4] Ivan Bjerre Damgård, *A design principle for hash functions*. Advances in Cryptology – CRYPTO '89 (G. Brassard, ed.), Lecture Notes in Computer Science 435, pp. 416–427. Springer, 1989.

[5] Aldo DeLuca and Stefano Varrichio, *Finiteness and Regularity in Semigroups and Formal Languages*. Springer, 1999.

[6] Robert Dilworth, *A decomposition theorem for partially ordered sets*, The Annals of Mathematics 51 (1950), 161–166.

[7] Hans Dobbertin, *Cryptanalysis of MD4*, Journal of Cryptology 11 (1998), 253–271.

[8] Kimmo Halunen, Juha Kortelainen and Tuomas Kortelainen, *Combinatorial Multicollision Attacks on Generalized Iterated Hash Functions*. Eighth Australasian Information Security Conference (AISC 2010) (C. Boyd and W. Susilo, eds.), CRPIT 105, pp. 86–93. ACS, Brisbane, Australia, 2010.

[9] Jonathan J. Hoch and Adi Shamir, *Breaking the ICE – Finding multicollisions in iterated concatenated and expanded (ICE) hash functions*. Fast Software Encryption – FSE '06 (Matthew J. B. Robshaw, ed.), Lecture Notes in Computer Science 4047, pp. 179–194. Springer, 2006.

[10] Antoine Joux, *Multicollisions in iterated hash functions. Application to cascaded constructions*. Advances in Cryptology – CRYPTO '04 (Matthew K. Franklin, ed.), Lecture Notes in Computer Science 3152, pp. 306–316. Springer, 2004.

[11] John Kelsey and Tadayoshi Kohno, *Herding hash functions and the Nostradamus attack*. Advances in Cryptology – EUROCRYPT '06 (Serge Vaudenay, ed.), Lecture Notes in Computer Science 4004, pp. 183–200. Springer, 2006.

[12] Vlastimil Klima, *Finding MD5 collisions on a notebook PC using multi-message modifications*, Cryptology ePrint Archive, Report 2005/102, 2005, `http://eprint.iacr.org/2005/102`.

[13] Vlastimil Klima, *Huge multicollisions and multipreimages of hash functions BLENDER-n*, Cryptology ePrint Archive, Report 2009/006, 2009, `http://eprint.iacr.org/2009/006`.

[14] Alfred J. Menezes, Paul C. van Oorschot and Scott A. Vanstone (eds.), *Handbook of Applied Cryptography*. CRC Press, 1996.

[15] Ralph C. Merkle, *One Way Hash Functions and DES*. Advances in Cryptology – CRYPTO '89 (G. Brassard, ed.), Lecture Notes in Computer Science 435, pp. 428–446. Springer-Verlag, 1990.

[16] Mridul Nandi and Douglas R. Stinson, *Multicollision attacks on some generalized sequential hash functions*, IEEE Transactions on Information Theory 53 (2007), 759–767.

[17] Marc Stevens, *Fast collision attack on MD5*, Cryptology ePrint Archive, Report 2006/104, 2006, `http://eprint.iacr.org/2006/104`.

[18] Kazuhiro Suzuki, Dongvu Tonien, Kaoru Kurosawa and Koji Toyota, *Birthday paradox for multi-collisions*, IEICE Transactions 91-A (2008), 39–45.

[19] Xiaoyun Wang, Yiqun Lisa Yin and Hongbo Yu, *Finding collisions in the full SHA-1*. Advances in Cryptology – CRYPTO '05 (Victor Shoup, ed.), Lecture Notes in Computer Science 3621, pp. 17–36. Springer, 2005.

[20] Xiaoyun Wang and Hongbo Yu, *How to break MD5 and other hash functions*. Advances in Cryptology – EUROCRYPT '05 (Ronald Cramer, ed.), Lecture Notes in Computer Science 3494, pp. 19–35. Springer, 2005.

[21] Hongbo Yu and Xiaoyun Wang, *Multi-collision attack on the compression functions of MD4 and 3-pass HAVAL*. Information Security and Cryptology – ICISC '07 (Kil-Hyun Nam and Gwangsoo Rhee, eds.), Lecture Notes in Computer Science 4817, pp. 206–226. Springer, 2007.

Received August 30, 2009; revised June 10, 2010; accepted September 6, 2010.

Author information

Juha Kortelainen, Department of Information Processing Science, University of Oulu, Finland.
E-mail: `juha.kortelainen@oulu.fi`

Kimmo Halunen, Secure Programming Group, Department of Electrical and Information Engineering, University of Oulu, Finland.
E-mail: khalunen@ee.oulu.fi

Tuomas Kortelainen, Mathematics Division, Department of Electrical and Information Engineering, University of Oulu, Finland.
E-mail: tuomas.kortelainen@oulu.fi