# Combinatorics on words in information security: Unavoidable regularities in the construction of multicollision attacks on iterated hash functions

Juha Kortelainen
Department of Information Processing Science, University of Oulu, Finland
jkortela@tols16.oulu.fi

Classically in combinatorics on words one studies unavoidable regularities that appear in sufficiently long strings of symbols over a fixed size alphabet. In this paper we take another viewpoint and focus on combinatorial properties of long words in which the number of occurrences of any symbol is restritced by a fixed constant. We then demonstrate the connection of these properties to constructing multicollision attacks on so called generalized iterated hash functions.

## 1 Introduction

In combinatorics on words, the theory of 'unavoidable regularities' usually concerns properties of long words over a fixed finite alphabet. Famous classical results in general combinatorics and algebra such as theorems of Ramsey, Shirshov and van der Waerden can then be straightforwardly exploited ([2], [9], [11], [12], [13]). The theory can be applied in the study of finiteness conditions for semigroups and (through the concept of syntactic monoid) also in regular languages and finite automata. To give the reader a view of the traditional basic results in unavoidable regularities we list some of its most noteworthy achievements.

Ramsey's Theorem immediately implies

**Theorem 1 (Repeated Patterns [2])** *For all positive integers $m$ and $n$ there exists a positive integer $R(m,n)$ satisfying the following. Given an alphabet $A$ and a partition $\{A_i\}_{i=1}^m$ of $A^+$ into $m$ sets, if $w \in A^+$ is any word of length at least $R(m,n)$, then $w$ is in $A^*A_j^nA^*$ for some $j \in \{1,2,\ldots,m\}$.*

Let $A$ be an alphabet totally ordered by $<$. We extend the order $<$ to the *lexiographic order* $<_{lex}$ of $A^*$ as follows. For all $u,v \in A^*$: $u <_{lex} v$ if either $v \in uA^+$ or $u = xay$ and $v = xbz$ for some $x,y,z \in A^*$ and $a,b \in A$ for which $a < b$.

Given a positive integer $n$, the word $w \in A^*$ is *$n$-divided* if there exist words $u,x_1,x_2,\ldots,x_n,v$ in $A^*$ such that $w = ux_1x_2\cdots x_nv$ and

$$w <_{lex} ux_{\sigma(1)}x_{\sigma(2)}\cdots x_{\sigma(n)}v$$

for any nontrivial permutation $\sigma : \{1,2,\ldots,n\} \to \{1,2,\ldots,n\}$.

**Theorem 2 (Shirshov [8, 9, 12])** *Let $A$ be an alphabet of $k$ symbols and $p$ and $n$ positive integers with $p \geq 2n$. There then exists a positive integer $S(k,p,n)$ such that any word in $A^*$ of length at least $S(k,p,n)$ either is $n$-divided or contains a $p$th power of a nonempty word of length at most $n-1$.*

Let $w = a_1a_2\cdots a_m$ where $a_i \in A$ for $i = 1,2,\ldots,m$. A *cadence* of $w$ is any sequence $(i_1,i_2,\ldots,i_s)$ of integers such that

$$0 < i_1 < i_2 < \cdots < i_s \text{ and } a_{i_1} = a_{i_2} = \cdots = a_{i_s} .$$

P. Ambrož, Š. Holub and Z. Masáková (Eds.):
8th International Conference WORDS 2011
EPTCS 63, 2011, pp. 22–29, doi:10.4204/EPTCS.63.4

Here the number $s$ is the *order* of the cadence. The cadence $(i_1, i_2, \ldots, i_s)$ is *arithmetic* if there exists a positive integer $d$ such that $i_j = i_1 + (j-1)d$ for $j = 1, 2, \ldots, s$.

The celebrated van der Waerden's theorem can be reformulated in words as follows.

**Theorem 3 (van der Waerden [8, 9])** *Let $A$ be an alphabet of $k$ symbols and $s$ a positive integer. There then exists a positive integer $W(k,s)$ such that any word in $A^*$ of length at least $W(k,s)$ possesses an arithmetic cadence of order $s$.*

Combinatorial problems are also encountered in information security, for example, when designing and investigating hash functions, techniques used in message authentication and digital signature schemes. A *hash function of length $n$* (where $n \in \mathbb{N}_+$) is a mapping $\mathtt{H} : \{0,1\}^* \to \{0,1\}^n$. For computing resource reasons, practical hash functions are often *iterative*, i.e., they are based on some finite compression function and an initial hash value. For more details, see subsection 3.1.

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

There are three main security properties that usually are required from a hash function $\mathtt{H}$: *collision resistance*, *preimage resistance*, and *second preimage resistance*.

**Collision resistance:** It is computationally infeasible to find $x, x' \in \{0,1\}^*$, $x \neq x'$, such that $\mathtt{H}(x) = \mathtt{H}(x')$.

**Preimage resistance:** Given any $y \in \{0,1\}^n$, it is computationally infeasible to find $x \in \{0,1\}^*$ such that $\mathtt{H}(x) = y$.

**Second preimage resistance:** Given any $x \in \{0,1\}^*$, it is computationally infeasible to find $x' \in \{0,1\}^*$, $x \neq x'$, such that $\mathtt{H}(x) = \mathtt{H}(x')$.

If we want to consider the resistance properties mathematically, the concept 'computationally infeasible' should be rigorously defined. Then the security of $H$ is compared to the security of a random oracle.

We thus say that $H$ is collision resistant (or possesses the collision resistance property) if to find $x, x' \in \{0,1\}^*$, $x \neq x'$, such that $\mathtt{H}(x) = \mathtt{H}(x')$ is (approximately) as difficult as to find $z, z' \in \{0,1\}^*$, $z \neq z'$, such that $\mathtt{G}(z) = \mathtt{G}'(z')$ for any random oracle hash function $\mathtt{G}$ of length $n$.

The concepts of preimage resistance and second preimage resistance can be defined analogously.

Given a set $C \subseteq \{0,1\}^*$ of finite cardinality $k > 1$, we say that $C$ is an *$k$-collision on* $\mathtt{H}$ if $\mathtt{H}(x) = \mathtt{H}(x')$ for all $x, x' \in C$. Any 2-collison is also called a collision (on $\mathtt{H}$).

The sharpened definitions allow us to define a fourth security property, the so called multicollision resistance: The hash function $\mathtt{H}$ is *multicollision resistant* if, for each $k \in \mathbb{N}_+$, to find an $k$-collison on $\mathtt{H}$ is (approximately) as difficult as to find an $k$-collison on any random oracle hash function $\mathtt{G}$ of length $n$.

Our conciderations are connected to multicollison resistance. Given a message $x = x_1 x_2 \cdots x_l$ where $x_1, x_2, \ldots, x_l$ are the (equally long) blocks of $x$, the value of a generalized iterated hash function on $x$ is based on the values of a finite compression function on the message blocks $x_1, x_2, \ldots, x_l$. A nonempty word $\alpha$ over the alphabet $\{1, 2, \ldots, l\}$ may then tell us in which order and how many times each block $x_i$ is expended by the compression function when producing the value of the respective generalized iterated hash function. Since the length of messages vary, we get to consider sequences of words $\alpha_1, \alpha_2, \ldots$ in which, for each $l \in \{1, 2, \ldots\}$, the word $\alpha_l \in \{1, 2, \ldots, l\}^*$ is related to messages with $l$ blocks. Practical applications state one more limitation: given a message of any length, a fixed block is to be consumed by the compression function only a restricted number ($q$, say) of times when computing the generalized iterated hash function value. Thus in the sequence $\alpha_1, \alpha_2, \ldots$ we assume that for each $l \in \{1, 2, \ldots\}$ and $m \in \{1, 2, \ldots, l\}$, the number $|\alpha_l|_m$ of occurrences of the symbol $m$ in the word $\alpha_l$ is at most $q$.

What can be said about the general combinatorial properties of the word $\alpha_l$ when $l$ grows? More generally: which kind of unavoidable regularities appear in sufficiently long words in which the number of occurrences of any symbol is bounded by a fixed constant?

As is easy to imagine, the regularities in the words $\alpha_l$ weaken the respective generalized iterated hash function against multicollision attacks. This topic was first studied in [3], see also [4, 10, 1, 6, 7, 5]. We shall present combinatorial results on words which imply that $q$-bounded generalized iterated hash functions are not multicollision resistant.

We proceed in the following order. In the next section basic concepts are briefly given. In the third section we first introduce the basics of generalized iterated hash functions. The connection to combinatorics on words is then established. The fourth section contains the necessary combinatorial results. Finally, the last section contains conclusions and further research proposals.

## 2 Preliminaries

Let $\mathbb{N} = \{0,1,2,\ldots\}$ be the set of all natural numbers and $\mathbb{N}_+ = \mathbb{N}\setminus\{0\}$. For each finite set $S$, let $|S|$ be the *cardinality* of $S$ that is to say, the number of elements in $S$.

Let $A$ be a finite alphabet and $\alpha \in A^+$. The length of the word $\alpha$ is denoted by $|\alpha|$; for each $a \in A$, let $|\alpha|_a$ be the number of occurrences of the letter $a$ in $\alpha$, and let $\mathrm{alph}(\alpha)$ denote the set of all letters occurring in $\alpha$ at least once. The empty word is denoted by $\varepsilon$. A permutation of $A$ is any word $\beta \in A^+$ such that $|\beta|_a = 1$ for each $a \in A$.

Let $B \subseteq A$. Then the *projection morphism* from $A^*$ into $B^*$, denoted by $\Pi^A_B$ is defined by $\Pi^A_B(b) = b$ if $b \in B$ and $\Pi^A_B(b) = \varepsilon$ if $b \in A\setminus B$. We write $\Pi_B$ instead of $\Pi^A_B$ when $A$ is understood. Define the word $(\alpha)_B$ as follows: $(\alpha)_B = \varepsilon$ if $\pi_B(\alpha) = \varepsilon$ and $(\alpha)_B = a_1a_2\cdots a_s$ if $\pi_B(\alpha) \in a_1^+a_2^+\cdots a_s^+$, where $s \in \mathbb{N}_+$, $a_1,a_2,\ldots,a_s \in B$, and $a_i \neq a_{i+1}$ for $i = 1,2,\ldots,s-1$.

## 3 Hash functions and collisions

In this section we first present a compact lead-in to (generalized) iterated hash functions. Later we wish to point out how certain results in combinatorics on words are interconnected to successful multicollision construction on these type of hash functions.

### 3.1 Introduction to (generalized) iterated hash functions

Let $m,n \in \mathbb{N}_+$ be such that $m > n$. Then $H = \{0,1\}^n$ is the set of *hash values* (of length $n$) and $B = \{0,1\}^m$) is the set of *message blocks* (of length $m$). Any $w \in B^+$ is a *message*. Given a mapping $f : H\times B \to H$, call $f$ a *compression function* (of length $n$ and block size $m$).

Define the function $f^+ : H\times B^+ \to H$ inductively as follows. For each $h \in H$, $b \in B$ and $x \in B^+$, let $f^+(h,b) = f(h,b)$ and $f^+(h,bx) = f^+(f(h,b),x)$. Note that $f^+$ is nothing but an iterative generalization of the compression function $f$.

Let $l \in \mathbb{N}_+$ and $\alpha$ be a nonemptyword such that $\mathrm{alph}(\alpha) \subseteq \mathbb{N}_l$. Then $\alpha = i_1i_2\cdots i_s$, where $s \in \mathbb{N}_+$ and $i_j \in \mathbb{N}_l$ for $j = 1,2,\ldots,s$. Define the *iterated compression function* $f_\alpha : H\times B^l \to H$ (based on $\alpha$ and $f$) by

$$f_\alpha(h,b_1b_2\cdots b_l) \;=\; f^+(h,b_{i_1}b_{i_2}\cdots b_{i_s})$$

for each $h \in H$ and $b_1, b_2, \ldots, b_l \in B$. Note that clearly $\alpha$ only declares how many times and in which order the message blocks $b_1, b_2, \ldots, b_l$ are used when creating the (hash) value $f_\alpha(h, b_1 b_2 \ldots b_l)$ of the message $b_1 b_2 \cdots b_l$.

Given $k \in \mathbb{N}_+$ and $h_0 \in H$, a *$k$-collision* (*with initial value $h_0$*) *in the iterated compression function $f_\alpha$* is a set $C \subseteq B^l$ such that the following holds:

1. The cardinality of $C$ is $k$;
2. For all $u, v \in C$ we have $f_\alpha(h_0, u) = f_\alpha(h_0, v)$; and
3. For any pair of distinct messages $u = u_1 u_2 \cdots u_l$ and $v = v_1 v_2 \cdots v_l$ in $C$ such that $u_i, v_i \in B$ for $i = 1, 2, \ldots, l$, there exists $j \in \{1, 2, \ldots, l\}$ for which $u_j \neq v_j$.

For each $j \in \mathbb{N}_+$, let now $\alpha_j \in \mathbb{N}_j^+$ be such that $\mathrm{alph}(\alpha_j) = \mathbb{N}_j$. Denote $\hat{\alpha} = (\alpha_1, \alpha_2, \ldots)$. Define the *generalized iterated hash function* (a gihf for short) $\mathtt{H}_{\hat{\alpha},f} : H \times B^+ \to H$ (based on $\hat{\alpha}$ and $f$) as follows: Given the initial value $h_0 \in H$ and the message $x \in B^j$, $j \in \mathbb{N}_+$, let

$$\mathtt{H}_{\hat{\alpha},f}(h_0, x) = f_{\alpha_j}(h_0, x) \ .$$

Thus, given any message $x$ of $j$ blocks and hash value $h_0$, to obtain the value $\mathtt{H}_{\hat{\alpha},f}(h_0, x)$, we just pick the word $\alpha_j$ from the sequence $\hat{\alpha}$ and compute $f_{\alpha_j}(h_0, x)$. For more details, see [6] and [3].

**Rermark 1** *A traditional iterated hash function $\mathtt{H} : B^+ \to H$ based on $f$ (with initial value $h_0 \in H$) can of course be defined by $\mathtt{H}(u) = f^+(h_0, u)$ for each $u \in B^+$. On the other hand $\mathtt{H}$ is a generalized iterated hash function $\mathtt{H}_{\hat{\alpha},f} : H \times B^+ \to H$ 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 $h_0$. Note that almost all hash functions used nowadays in practise are of this form.*

Given $k \in \mathbb{N}_+$ and $h_0 \in H$, a *$k$-collision in the generalized iterated hash function $\mathtt{H}_{\hat{\alpha},f}$* (*with initial value $h_0$*) is a set $C$ of $k$ messages such that for all $u, v \in C$, $|u| = |v|$ and $\mathtt{H}_{\hat{\alpha},f}(h_0, u) = \mathtt{H}_{\hat{\alpha},f}(h_0, v)$. Now suppose that $C$ is a $k$-collision in $\mathtt{H}_{\hat{\alpha},f}$ with initial value $h_0$. Let $l \in \mathbb{N}_+$ be such that $C \subseteq B^l$, i.e., the length in blocks of each message in $C$ is $l$. Then, by definition, for each $u, v \in C$, the equality $f_{\alpha_l}(h_0, u) = f_{\alpha_l}(h_0, v)$ holds. Since $\mathrm{alph}(\alpha_l) = \mathbb{N}_l$ (and thus each symbol in $\mathbb{N}_l$ occurs in $\mathrm{alph}(\alpha)$), the set $C$ is a $k$-collision in $f_{\alpha_l}$ with initial value $h_0$. Thus, a $k$-collision in the generalized iterated hash function $\mathtt{H}_{\hat{\alpha},f}$ necessarily by definition, is a $k$-collision in the iterated compression function $f_{\alpha_l}$ for some $l \in \mathbb{N}_+$.

Now, in our security model, the *attacker* tries to find a $k$-collision in $\mathtt{H}_{\hat{\alpha},f}$. We assume that the attacker knows how $\mathtt{H}_{\hat{\alpha},f}$ 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 $(h, b) \in H \times B$) on $f$ and get the respective *responses* (values $f(h, b) \in H$).

We thus define the (*message*) *complexity of a $k$-collision* in $\mathtt{H}_{\hat{\alpha},f}$ to be the expected number of queries on the compression function $f$ that is needed to create a multicollision of size $k$ in $\mathtt{H}_{\hat{\alpha},f}$ with any initial value $h \in H$.

According to the (generalized) *birthday paradox*, a $k$-collision for any compression function $f$ of length $n$ can be found (with probability approx. $\frac{1}{2}$) by hashing $(k!)^{\frac{1}{k}} 2^{\frac{n(k-1)}{k}}$ messages [14] if we assume that there is no memory restrictions. Two remarks can be made immediately:

- In the case $k = 2$ approximately $\sqrt{2} \cdot 2^{\frac{n}{2}}$ hashings (queries on $f$) are needed; intuitively many of us would expect the number to be around $2^{n-1}$.
- For each $k$ in $\mathbb{N}_+$, finding a $(k+1)$-collision consumes much more resources than finding a $k$-collision.

Of course, when attacking, for instance, against an iterated hash function based on a random oracle compression function of length $n$, the attacker needs a lot of computing power when $n$ is large; to create a 2-collison requires approximately $\sqrt{2}\cdot 2^{\frac{n}{2}}$ queries on $f$ and this is resource consuming.

The paper [4] presents a clever way to find a $2^r$-collision in the traditional iterated hash function $\mathtt{H}$ (see Remark 1) for any $r \in \mathbb{N}_+$. The attacker starts from the initial value $h_0$ and searches two distinct message blocks $b_1$, $b_1'$ such that $f(h_0,b_1) = f(h_0,b_1')$ and denotes $h_1 = f(h_0,b_1)$. By the birthday paradox, the expected number of queries on $f$ is $\tilde{a}\,2^{\frac{n}{2}}$ , where $\tilde{a}$ is approximately 2.5. Then, for each $i = 2,3,\ldots,r-1$, the attacker continues by searching message blocks $b_i$ and $b_i'$ such that $b_i \neq b_i'$ and $f(h_{i-1},b_i) = f(h_{i-1},b_i')$ and and stating $h_i = f(h_{i-1},b_i)$. Now the set $C = \{b_1,b_1'\}\times\{b_2,b_2'\}\times\cdots\times\{b_r,b_r'\}$ is $2^r$-collision in $\mathtt{H}$. The expected number of queries on $f$ is clearly $\tilde{a}\,r2^{\frac{n}{2}}$, i.e., the work the attacker is expected to do is only $r$ times greater than the work she or he has to do to find a single 2-collision. The size of the multicollision grows exponentially while the need of resources increases linearly.

The question arises whether or not the ideas of Joux can be applied in a more broad setting, i.e., can Joux's approach be used to multicollisions in certain generalized iterated hash functions?

In the following we shall see that this indeed is possible. 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 $i \in \mathbb{N}_j$. The gihf $\mathtt{H}_{\hat{\alpha},f}$ is *$q$-bounded* if $\hat{\alpha}$ is $q$-bounded. Note that Joux's method is easy to apply to any 1-bounded generalized iterated hash function.

Is it possible to extend Joux's method furthermore to be adapted to $q$-bounded gihfs, when $q > 1$? This question has been investigated first for 2-bounded gihfs in [10] and then for any $q$-bounded gihf in [3] (see also [6]). It turned out that it is possible to create $2^r$-collision in any $q$-bounded gihf with $O(g(n,q,r)\,2^{\frac{n}{2}})$ queries on $f$, where $g(n,q,r)$ is function of $n,q$ and $r$ which is polynomial with respect to $n$ and $r$ but double exponential with respect to $q$.

The idea behind the successful construction of the attack is the fact that since $\hat{\alpha}$ is $q$-bounded, unavoidable regularities start to appear in the word $\alpha_l$ of $\hat{\alpha}$ when $l$ is increased. More accurately, choosing $l$ large enough, yet so that $|\mathrm{alph}(\alpha_l)|$ depends only polynomially on $n$ and $r$ (albeit double exponentially in $q$), a number $p \in \{1,2,\ldots,q\}$ and a set $A \subseteq \mathrm{alph}(\alpha_l)$ of cardinality $|A| = n^{p-1}r$ can be found such that

(P1) $\alpha_l = \beta_1\beta_2\cdots\beta_p$ the word $(\beta_i)_A$ is a permutation of $A$ for $i = 1,2,\ldots,p$; and

(P2) for any $i \in \{1,2,\ldots,p-1\}$, if $(\beta_i)_A = z_1z_2\cdots z_{n^{p-i}r}$ is a factorization of $(\beta_i)_A$ such that $|\mathrm{alph}(z_j)| = n^{i-1}$ for $j = 1,2,\ldots n^{p-i}r$ and $(\beta_{i+1})_A = u_1u_2\cdots u_{n^{p-i+1}r}$ is a factorization of $(\beta_{i+1})_A$ such that $|\mathrm{alph}(u_j)| = n^i$ for $j = 1,2,\ldots n^{p-i+1}r$, then for each $j_1 \in \{1,2,\ldots,n^{p-i}r\}$, there exists $j_2 \in \{1,2,\ldots,n^{p-i-1}r\}$ such that $\mathrm{alph}(z_{j_1}) \subseteq \mathrm{alph}(u_{j_2})$.

The property (P1) allows the attacker construct a $2^{|A|}$-collision $C_1$ in $f_{\beta_1}$ with any initial value $h_0$ so that the expected number of queries on $f$ is $\tilde{a}(|\beta_1|\,2^{\frac{n}{2}})$. The property (P2) ensures that based on the multicollision guaranteed by (P1), the attacker can proceed and, for $i = 2,3,\ldots,p$, create a $2^{n^{p-i}r}$-collision $C_i$ in $f_{\beta_1\beta_2\cdots\beta_i}$ so that the expected number of queries on $f$ is $\tilde{a}|\beta_1\beta_2\cdots\beta_i|\,2^{\frac{n}{2}}$. Thus finally a $2^r$-collision of complexity $\tilde{a}|\alpha|\,2^{\frac{n}{2}}$ in $\mathtt{H}_{\hat{\alpha},f}$ is generated.

Finally on the basis of the previous attack construction and (the future) Theorem 8, the following can be proved ([5]).

**Theorem 4** *Let $m$, $n$ and $q$ be positive integers such that $m > n$ and $q > 1$, $f : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}^n$ a compression function, and $\hat{\alpha} = (\alpha_1,\alpha_2,\ldots)$ a $q$-bounded sequence of words such that* $\mathrm{alph}(\alpha_l) = \mathbb{N}_l$ *for each $l \in \mathbb{N}_+$. Then, for each $r \in \mathbb{N}_+$, there exists a $2^r$-collision attack on the generalized iterated hash function $\mathtt{H}_{\hat{\alpha},f}$ such that the expected number of queries on $f$ is at most $\tilde{a}\,q\,N(n^{(q-1)^2}r^{2q-3},q)\,2^{\frac{n}{2}}$.*

**Rermark 2** *The inequality $N(m,q) < m^{2^{q-1}}$ (see Theorem 5) implies that*

$$N(n^{(q-1)^2}r^{2q-3},q) \;<\; n^{(q-1)^2 2^{q-1}}r^{(2q-3)\,2^{q-1}} \;.$$

The results in [14] imply that, given a random oracle hash function $\mathtt{G}$ of length $2^n$, the expected number of queries on $\mathtt{G}$ to find a $2^r$-collision is in $\Omega(2^{n\frac{2^r-1}{2^r}})$.

Call a generalized iterated hash function bounded if it is $q$-bounded for some $q \in \mathbb{N}_+$.

**Corollary 1** *There does not exist a bounded generalized iterated hash function that is multicollision resistant.*

### 3.2 Essential combinatorial results

We state a list of combinatorial results that imply Theorem 4. The main result in stated is the form of classical combinatorial theorems. For a proof, see [5].

**Theorem 5** *For all positive integers $m$ and $q$ there exists a (minimal) positive integer $N(m,q)$ such that if $\alpha$ is a word for which $|\mathrm{alph}(\alpha)| \geq N(m,q)$ and $|\alpha|_a \leq q$ for each $a \in \mathrm{alph}(\alpha)$, there exist $A \subseteq \mathrm{alph}(\alpha)$ with $|A| = m$, 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 $A$. Moreover, for all $m,q \in \mathbb{N}_+$ we have $N(m,q+1) \leq N(m^2-m+1,q)$.*

**Rermark 3** *Let $m \in \mathbb{N}_+$. In the case $q = 2$, the previous theorem gives us the boundary value $N(m,2) = m^2-m+1$. Let*

$$A = \{a_{i,j} | i = 1,2,\ldots,m-1,\ j = 1,2,\ldots,m\}$$

*be an alphabet of $m(m-1)$ symbols. Let furthermore*

$$\gamma_i = a_{i,1}a_{i,2}\cdots a_{i,m-1}a_{i,m}a_{i,m-1}a_{i,m-2}\cdots a_{i,1}$$

*for $i = 1,2,\ldots,m-1$ and $\alpha = \gamma_1\gamma_2\cdots\gamma_{m-1}$. It is quite straightforward to see that there does not exist an $m$-letter subalphabet of $A$ such that either (i) $(\alpha)_A$ is a permutation of $A$ or (ii) there exists a factorization $\alpha = \alpha_1\alpha_2$ such that $(\alpha_1)_A$ and $(\alpha_2)_A$ are both permutations of $A$. Thus $N(m,2) = m^2-m+1$ for $m \in \mathbb{N}_+$.*

Suppose now that $A$ and $\alpha = \alpha_1\alpha_2\cdots\alpha_p$ are as in Theorem 5, i.e., for all $i \in \{1,2,\ldots,p\}$, the word $(\alpha_i)_A$ is a permutation of $A$. To make our multicollision attack succeed, this is not yet sufficient. We need permutations $\beta_1$, $\beta_2$, ..., $\beta_p$ of an sufficiently large alphabet $B$ such that when factoring $\beta_i = \beta_{i1}\beta_{i2}\cdots\beta_{id_i}$ into $d_i \in \mathbb{N}_+$ equal length factors for $i = 1,2,\ldots,p$ where $d_j$ divides $d_{j+1}$ and the following holds: for each $i \in \{1,2,\ldots,p-1\}$ and $j_1 \in \{1,2,\ldots,d_i\}$ there exists $j_2 \in \{1,2,\ldots,d_{i+1}\}$ such that $\mathrm{alph}(\beta_{ij_1}) \subseteq \mathrm{alph}(\beta_{i+1,j_2})$. Only then we can, starting from the first permutation (and the word $\alpha_1$) roll on our attack well. Above the permutations $\beta_1,\beta_2,\ldots,\beta_p$ are induced by the words $\alpha_1,\alpha_2,\ldots,\alpha_p$, respectively, when $\alpha$ is long enough (or equivalently, the alphabet $\mathrm{alph}((\alpha)$ is sufficiently large). That these permutations always can be found, is verified in the following three combinatorial results.

We wish to further study the mutual structure of permutations in long words guaranteed by Theorem 5. By increasing the length of the word $\alpha$ the permutations are forced to possess certain stronger structural properties. The motives are, besides our interest in combinatorics on words, in information security applications. The connection of the results to creating multicollisions on generalized iterated hash functions is more accurately, albeit informally, described in Section 5.

As the first step of our reasoning we need an application of the famous Hall's Matching Theorem. For the proof, see [6] and [3].

**Theorem 6 (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 $i = 1,2,\ldots,k$.*

The next theorem is also from [6]. It is an inductive generalization of Partition Theorem to different size of factorizations. For the proof, see [6].

**Theorem 7 (Factorization Theorem)** *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 $w_{r+1} = u_1 u_2 \cdots u_{d_r}$ is the factorization $w_{r+1}$ into $d_r$ equal length $(= d_0 d_1^2 d_2^2 \cdots d_{r-1}^2 d_r)$ blocks, then $\pi_B(w_{r+1}) = \pi_B(u_1)\pi_B(u_2) \cdots \pi_B(u_{d_r})$ is the factorization of $\pi_B(w_{r+1})$ into $d_r$ equal length $(= \frac{d_0}{d_r})$ blocks.*

In fact what we need in our considerations is the following

**Corollary 2** *Let $d_0, d$ and $r$ be positive integers such that $d$ divides $d_0$, $A$ an alphabet of cardinality $|A| = d_0 d^{2r}$, and $w_1, w_2, \ldots, w_{r+1}$ permutations of $A$. Then there exists a subset $B$ of $A$ of cardinality $|B| = d_0$ satisfying the following. Let $p, q \in \{1, 2, \ldots, r+1\}$ and $\pi_B(w_p) = x_1 x_2 \cdots x_d$ the factorization of $\pi_B(w_p)$ and $\pi_B(w_q) = y_1 y_2 \cdots y_d$ the factorization of $\pi_B(w_q)$ into $d$ equal length $(= \frac{d_0}{d})$ blocks, then for each $i \in \{1, 2, \ldots, d\}$, there exists $j \in \{1, 2, \ldots, d\}$ such that $\mathrm{alph}(x_i) = \mathrm{alph}(y_j)$.*

The last result of this section combines the main result of this section (Theorem 5) to the previous combinatorial accomplishments. Theorem 8 is indispensable for the attack constrution in the end of Section 3.1.

**Theorem 8** *Let $\alpha$ be a word and $k \geq 2$, $n \geq 1$, and $q \geq 2$ integers such that*

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

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

*Then there exists $B \subseteq \mathrm{alph}(\alpha)$, $p \in \{1, 2, \ldots, q\}$ 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 $(\alpha_i)_B$ is a permutation of $B$ 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 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 $\mathrm{alph}(z_{j_1}) \subseteq \mathrm{alph}(u_{j_2})$.*

## 4 Conclusion

We have considered combinatorics on words from a fresh viewpoint which is induced by applications in information security. Some small steps have already been taken in the new research frame. The results have been promising; they imply more efficient attacks on generalized iterated hash functions and, from their part, confirm the fact that the iterative structure possesses certain generic security weaknesses.

*Research Problem*. Consider Theorem 5. The exact value of $N(m,q)$ is known only in the cases $m = 1$, $q = 1$ and $q = 2$: Trivially $N(1,q) = 1$ and $N(m,1) = m$, furthermore $N(m,2) = m^2 - m + 1$ (see Remark 3). It is probable that in general the number $N(m,q+1)$ is significantly smaller than $N(m^2 - m + 1, q)$. Moreover, we have not evaluated $N(m,q)$ from below at all. Find reasonable lower and upper bounds to $N(m,q)$ for $m > 1, q > 2$.

## References


[1] K. Halunen, J. Kortelainen & Kortelainen T. (2009): *Multicollision Attacks on Generalized Iterated Hash Functions*. In: *Eight Australasian Information Security Conference (AISC2010), Australian Computer Science Communications* 32, pp. 85–93.

[2] Michael A. Harrison (1978): *Introduction to formal language theory*. Addison-Wesley Publishing Co., Reading, Mass.

[3] Jonathan Hoch & Adi Shamir (2006): *Breaking the ICE - Finding Multicollisions in Iterated Concatenated and Expanded (ICE) Hash Functions*. In Matthew Robshaw, editor: *Fast Software Encryption*, *Lecture Notes in Computer Science* 4047, Springer Berlin / Heidelberg, pp. 179–194, doi:`10.1007/11799313_12`.

[4] Antoine Joux (2004): *Multicollisions in Iterated Hash Functions. Application to Cascaded Constructions*. In Matt Franklin, editor: *Advances in Cryptology – CRYPTO 2004*, *Lecture Notes in Computer Science* 3152, Springer Berlin / Heidelberg, pp. 99–213, doi:`10.1007/978-3-540-28628-8_19`.

[5] J. Kortelainen, T. Kortelainen & A. Vesanen (2011): *Unavoidable regularities in long words with bounded number of symbol occurrences*. Accepted to The 17th Annual International Computing and Combinatorics Conference (COCOON 2011).

[6] Juha Kortelainen, Kimmo Halunen & Tuomas Kortelainen (2010): *Multicollision attacks and generalized iterated hash functions*. *J. Math. Cryptol.* 4(3), pp. 239–270, doi:`10.1515/JMC.2010.010`.

[7] T. Kortelainen, J. Kortelainen & J. Halunen (2010): *Variants of multicollision attacks on iterated hash functions*. In: *6th China International Conference on Information Security and Cryptology (Inscrypt 2010)*, pp. 159–175.

[8] M. Lothaire (1997): *Combinatorics on words*. Cambridge Mathematical Library, Cambridge University Press, Cambridge, doi:`10.1017/CBO9780511566097`.

[9] Aldo de Luca & Stefano Varricchio (1999): *Finiteness and regularity in semigroups and formal languages*. Monographs in Theoretical Computer Science. An EATCS Series, Springer-Verlag, Berlin.

[10] Mridul Nandi & Douglas R. Stinson (2007): *Multicollision attacks on some generalized sequential hash functions*. *IEEE Trans. Inform. Theory* 53(2), pp. 759–767, doi:`10.1109/TIT.2006.889721`.

[11] A. Restivo (1977): *Mots sans répétitions et langages rationnels bornés*. *RAIRO Informat. Théor.* 11(3), pp. 197–202, iv.

[12] Antonio Restivo & Christophe Reutenauer (1983): *Some applications of a theorem of Shirshov to language theory*. *Inform. and Control* 57(2-3), pp. 205–213, doi:`10.1016/S0019-9958(83)80044-8`.

[13] Antonio Restivo & Christophe Reutenauer (1985): *Rational languages and the Burnside problem*. *Theoret. Comput. Sci.* 40(1), pp. 13–30, doi:`10.1016/0304-3975(85)90156-2`.

[14] Kazuhiro Suzuki, Dongvu Tonien, Kaoru Kurosawa & Koji Toyota (2006): *Birthday paradox for multicollisions*. In: *Information security and cryptology—ICISC 2006*, *Lecture Notes in Comput. Sci.* 4296, Springer, Berlin, pp. 29–40, doi:`10.1007/11927587_5`.