ON THE NUMBER OF SQUARES IN A FINITE WORD

Srečko Brlek[1] and Shuo Li*[2]

[1,2]*LACIM, Université du Québec à Montréal, Montréal, Québec, Canada*
brlek.srecko@uqam.ca, shuo.li.ismin@gmail.com
[2]*Departement of Mathematics and Statistics, The University of Winnipeg, Winnipeg, Manitoba, Canada*
sh.li@uwinnipeg.ca

Submitted: Apr 26, 2022; Accepted: Aug 14, 2024; Published: Mar 15, 2025

Abstract. Let u be a nonempty finite word, a *square* is a word of the form uu. In this paper, we prove that for a given finite word w, the number of distinct square factors of w is bounded by $|w| - |\operatorname{Alph}(w)|$, where $|w|$ denotes the length of w and $|\operatorname{Alph}(w)|$ denotes the number of distinct letters in w. This result answers positively a conjecture stated by Fraenkel and Simpson in 1998 and the d-step conjecture stated by Deza, Franek and Jiang in 2011.

Keywords. Combinatorics on words, squares, repetition

Mathematics Subject Classifications. 68R15, 68R10, 68R05

1. Introduction

The study of patterns in a word is related to several topics and it has various applications in several research domains, ranging from practical applications to theoretical considerations. Among these patterns, we can mention palindromes and repetitions such as squares, cubes, periods, overlaps, etc. They appear in different contexts in computer science such as data compression, searching algorithms, structure of indexes [Lot05], digital geometry [BGL13], in number theory about Diophantine approximation and transcendence statements [AA07], and in physics in connection with Schrödinger operators [ABCD03], among several others.

In this note, we are particularly interested in counting the number of distinct squares in a finite word. Fraenkel and Simpson established in [FS98] that for any finite word w of length n, the number of distinct squares is upper bounded by $2n$ and conjectured that this number is upper bounded by n. Later, Ilie [Ili07] strengthened this upper bound to $2n - \Theta(\log(n))$; Lam [Lam13] improved this result to $\frac{95}{48}n$; Deza, Franek and Thierry [DFT15] achieved an upper bound of $\frac{11}{6}n$; Thierry [Thi20] refined this bound to $\frac{3}{2}n$. On the other hand, Deza, Franek, and Jiang conjectured in [DFJ11] (see also [DF14]) that the number of distinct squares in a finite word of length n

*benefited from the support of a postdoctoral fellowship provided by the Laboratoire d'Algèbre, de Combinatoire et d'InformatiqueMathématique (LACIM) at the Université du Québec à Montréal.

with d distinct letters is upper bounded by $n-d$. It is a stronger version of the conjecture stated by Fraenkel and Simpson in [FS98]. Moreover, Deza et al. showed in [DFJ11] that the upper bound $n-d$ is optimal when n is small. In what follows, we prove the conjecture of Deza, et al.:

Theorem 1. *For any finite word w, the number $S(w)$ of its distinct square factors satisfies*

$$S(w) \leqslant |w| - |\mathrm{Alph}(w)|.$$

The strategy of the proof is the following: for a given word w, we first recall the definition of Rauzy graph and define *small circuits* in it; we then prove that the total number of small circuits in the union of the Rauzy graphs of w is bounded by $|w| - |\mathrm{Alph}(w)|$; finally we conclude by building an injection from the set of distinct squares of w into the set of small circuits in the union of the Rauzy graphs of w.

2. Preliminaries

Let us recall the basic terminology about words from Lothaire [Lot05]. A*word* is a finite sequence $w = w_1w_2\cdots w_n$ of *letters* or symbols. The *length* $|w|$ of w is n and w_i is the letter in *position* i. The *empty word* is of length 0 and is denoted by ε. The *concatenation of* $w = w_1w_2\cdots w_n$ *and* $v = v_1v_2\cdots v_m$ is defined as

$$wv = w_1w_2\cdots w_nv_1v_2\cdots v_m.$$

In particular, for any word u, we have $u = \varepsilon u = u\varepsilon$.

The *alphabet* $A = \mathrm{Alph}(w) = \{w_i \mid 1 \leqslant i \leqslant n\}$ is equipped with a total order $\prec$ which extends lexicographically to the set A^* of all words over A.

A word u is called a *factor* of w if $w = pus$ for some words p, s. The i-th prefix ending at position i is denoted $w_p(i) = w_1w_2\cdots w_i$ and the i-th suffix starting at position i is $w_s(i) = w_iw_{i+1}\cdots w_t$. Hence every word $w = w_1w_2\cdots w_n$ factorizes as $w = w_p(i-1)w_s(i)$. Of course, $w_s(1) = w$ and $w_p(0) = \varepsilon$.

The set of all length-i factors of w is denoted $\mathrm{Fac}_i(w)$ and the set of all factors of w is $\mathrm{Fac}(w)$.

Two finite words u and v are *conjugates* when there exist words x, y such that $u = xy$ and $v = yx$. The conjugacy class of a word w is denoted by $[w]$. Thus,

$$[w] = \{v | v = w_s(i)w_p(i-1), i = 1, 2, \ldots, n\}.$$

For any natural number k, the *k-power* of a nonempty finite word u is the concatenation of k copies of u, and it is denoted by u^k. In particular, a *square* is a word w of the form $w = uu$. A word w is said to be *primitive* if it is not a power greater than or equal to 2 of a word distinct from w. For any word u and any rational number $\alpha = \frac{m}{|w|}$, the α-power of u is defined to be u^au_0 where u_0 is a prefix of u, $a = \lfloor\alpha\rfloor$ is the integer part of α, and $|u^au_0| = m$. The α-power of u is denoted by u^α. A word w is said to be *of the period* n if there exists a word u and a positive rational number $\alpha \geqslant 1$ such that $|u| = n$ and $w = u^\alpha$, the period n is said to be *the smallest period* if for any prefix u' of w satisfying $|u'| < n$, w is not a (rational) power of u'.

For a primitive word p, the lexicographically least element in its conjugacy class $[p]$ is called a *Lyndon word*.

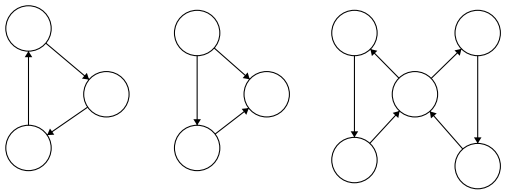


Figure 2.1: An elementary circuit, an elementary cycle and a nonelementary circuit.

Definition 2. *Let w be any finite word and let m be an integer such that $m \geqslant |w|$, then we define*

$$[w]_m = \left\{ p^{\frac{m}{|w|}} | p \in [w] \right\}.$$

Example 3. Let $w = aba$, then $[w] = [w]_3 = \{aba, aab, baa\}$, $[w]_4 = \{abaa, aaba, baab\}$ and $[w]_5 = \{abaab, aabaa, baaba\}$.

The following basic lemmas concerning repetitions will be useful :

Lemma 4 (Fine and Wilf [FW65])**.** *Let w be a word having periods k and l. If $|w| \geqslant k + l - \gcd(k, l)$ then $\gcd(k, l)$ is also a period of w.*

Lemma 5 (Lyndon and Schützenberger [LS62])**.** *Let x and y be two words such that $xy = yx$. Then there exist a primitive word p and two nonnegative integers i, j such that $x = p^i$ and $y = p^j$.*

Lemma 6 (Lyndon and Schützenberger [LS62])**.** *Let x, y, z be three words such that $xy = yz$ and $|y| \neq 0$. Then there exist two words u, v with $|u| \neq 0$ and some positive integer i such that $x = uv, y = (uv)^i u, z = vu$.*

Now, we recall some basic definitions and properties concerning graphs mainly from Berge [Ber82].

Let $G = (V, E)$ be a directed graph such that V is the set of its vertices and E is the set of its edges. A *chain* of *length* k is a sequence of edges $e_1, e_2, \ldots, e_k$, such that for each i satisfying $1 < i < k$, e_i has one vertex in common with the preceding edge e_{i-1} and another vertex in common with e_{i+1}. A *path* is a chain such that, for each i satisfying $1 \leqslant i < k$, the terminal vertex of e_i coincides with the initial vertex of e_{i+1}. A chain or a path is *closed* if it begins and ends at the same vertex. A *cycle* is a closed chain and *circuit* is a closed path. Obviously, any circuit is a cycle, but a cycle may not be a circuit (see Figure 2.1). A cycle $C = (V, E)$ (or a circuit) is *elementary* if each vertex on the chain occurs exactly twice: in this case the *length* of C is $|V| = |E|$.

A graph G is *weakly connected*, if for any pair of distinct vertices (v_1, v_2), there exists a chain with edges $e_1, e_2, \ldots, e_k$ such that v_1 is an edge of e_1 and v_2 is an edge of e_k. For a weakly connected graph G with l edges and s vertices, the number $\chi(G) = l - s + 1$ is the *cyclomatic number* of G.

Let G be a graph of l edges $e_1, e_2, \ldots, e_l$ and assign an orientation for each edge of G. For any cycle C in G, the *vector-cycle corresponding to* C is the vector $\mu(C) = (c_1, c_2, \ldots, c_l)$ in

the l-dimensional space $\mathbb{R}^l$ such that, for all i satisfying $1 \leqslant i \leqslant l$, $c_i = r_i - s_i$ if the edge e_i is included r_i times in the sense of its orientation and s_i times in the opposite sense in the cycle C. A set of cycles $C_1, C_2, \ldots, C_t$, is called *independent* if the corresponding vectors are linearly independent. Note that the independence of the cycles is invariant with respect to the selection of the default orientations.

Theorem 7 (Th. 2, Chap. 4 in [Ber82])**.** *The cyclomatic number of a graph is the maximum number of independent cycles in this graph.*

3. Graph of factors and small circuits

The graph of factors of a word is a convenient representation for our purpose, and was introduced by Rauzy [Rau83].

Definition 8. *Let w be a word of length n. For any integer i such that $1 \leqslant i \leqslant n$, the Rauzy graph $\Gamma_i(w)$ is a directed graph whose set of vertices is* $\mathrm{Fac}_i(w)$ *and the set of edges is* $\mathrm{Fac}_{i+1}(w)$*. An edge $e \in \mathrm{Fac}_{i+1}(w)$ starts at the vertex u and ends at the vertex v, if and only if u is a prefix and v is a suffix of e.*

For each i, it is a subgraph of the de Bruijn graph where the set of vertices is A^i [dB46]. Building the Rauzy graph is straightforward by sequentially reading factors of length i, and therefore $\Gamma_i(w)$ is a weakly connected graph for each i.

The family of graphs $\Gamma_i(w), 1 \leqslant i \leqslant n$ are disjoint, and for later use, we set

$$\Gamma(w) = \bigcup_{i=1}^{n} \Gamma_i(w).$$

Definition 9. *Let $\Gamma_i(w)$ be a Rauzy graph of w. An elementary circuit C in $\Gamma_i(w)$ is called* small *if the length of C is no larger than i.*

Let w be a word of length n. For each integer i, $1 \leqslant i \leqslant n$, let $sc_i(w)$ denote the number of small circuits in $\Gamma_i(w)$ and let $sc(w) = \sum_{i=1}^{n} sc_i(w)$ be the total number of small circuits in $\Gamma(w)$.

Lemma 10. *For any small circuit $C = (V, E)$ in $\Gamma_l(w)$, $1 \leqslant l \leqslant |w|$, there exists a Lyndon word q such that $|q| \leqslant l$, $V = [q]_l$ and $E = [q]_{l+1}$.*

Proof. Let $e_1, e_2, \ldots, e_k$ form the closed path of C and let $v_1, v_2, \ldots, v_k$ be the vertices of C such that e_i is from v_i to v_{i+1} for all i satisfying $1 \leqslant i < k$ and e_k is from v_k to v_1. From the definition of small circuits, $k \leqslant l$. Let p be a word of length k such that p_j is the last letter of e_j, $1 \leqslant j \leqslant k$. For each i satisfying $1 \leqslant i \leqslant k$, from the fact that the edges in the order of $e_i, e_{i+1}, \ldots, e_k, e_1, \ldots, e_{i-1}$ form a circuit, one can deduce that v_i is a suffix of $v_i q$, where q is a conjugate of p. Thus, there exists a word r of length k such that $v_i q = r v_i$. From Lemma 6, there exists two words u, t with $|u| \neq 0$ and a nonnegative integer j such that $r = ut, q = tu$ and $v_i = (ut)^j u$. Consequently, $v_i = r^{\frac{l}{k}}$. Further, as q and r are conjugates, r is also a conjugate

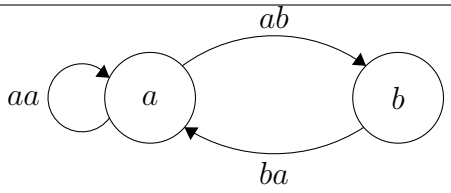
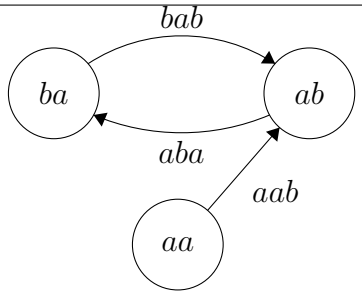
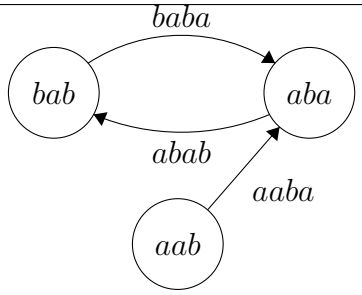
of p. Thus, the vertex set of C is $\left\{q^{\frac{l}{k}} | q \in [p]\right\}$. An analog argument for the edges shows that edge set of C is $\left\{q^{\frac{l+1}{k}} | q \in [p]\right\}$.

If p is not primitive, then there exists a primitive word s and an integer $m > 1$ such that $p = s^m$. In this case, $|[p]_{l+1}| = |[s^m]_{l+1}| = |[s]| < |p| = k$.

Let q be the Lyndon word in $[p]$, then $C = ([q]_l, [q]_{l+1})$. □

Notation 11. *To simplify the notation, denote* $C = ([q]_l, [q]_{l+1})$ *by* $C(q, l)$.

Example 12. For $u = aababa$ with the order induced by $a \prec b$, we give below $\Gamma_1(u), \Gamma_2(u)$ and $\Gamma_3(u)$ together with the small circuits therein:

$\Gamma_i(u)$	Rauzy Graphs	small circuits
$\Gamma_1(u)$		$C(a,1) = (\{a\}, \{aa\})$
$\Gamma_2(u)$		$C(ab,2) = (\{ab, ba\}, \{aba, bab\})$
$\Gamma_3(u)$		$C(ab,3) = (\{aba, bab\}, \{abab, baba\})$

Observe that, in $\Gamma_1(u)$, the circuit $(\{a, b\}, \{ab, ba\})$ is not small.

Definition 13. *For any small circuit, the* maximal edge *of this circuit is the lexicographically greatest element in its edge set.*

Lemma 14. *Each small circuit contains exactly one maximal edge. Further, in each Rauzy graph, the maximal edges of the small circuits are pairwise distinct.*

Proof. From Lemma 10 the edge set of any small circuit is of the form $\{p^t | p \in [q]\}$ for some fractional number $t > 1$ and some primitive word q. Thus, u^t is the lexicographically greatest

element in $\{p^t | p \in [q]\}$ if and only if u is the lexicographically greatest element in $[q]$. Hence the maximal edge of each small circuit is unique.

For the second part, we assume that v is the maximal edge of two distinct small circuits in the graph $\Gamma_m(w)$ for some word w and some integer m, $1 \leqslant m \leqslant |w|$. With the notation introduced in Notation 11, these two small circuits can be denoted respectively by $C(p, m)$ and $C(q, m)$, with p, q two distinct primitive words such that $|p| \leqslant m$, $|q| \leqslant m$. From Lemma 10, there exist two positive integers i and j and (possibly empty) prefixes p' and q' of p and q, such that $v = p^i p' = q^j q'$. First, it is clear that $|p'| \neq |q'|$. Otherwise, $p^i = q^j$. However, in this case, from Lemma 4, p and q cannot be both primitive. Without loss of generality, let us suppose $|p'| > |q'|$ and, consequently, $|p^i| < |q^j|$. Then there exists a prefix p'' of p such that $|p''| < |p|$ and $q^j = p^i p''$. However, p'' is also a prefix of p^i. Thus, there exists a word s such that $q^j = p''sp''$ and $p^i = p''s$.

Now we claim that v cannot be the maximal edge of $C(q, r)$. It is enough to prove that $q^j = p''sp''$ cannot be the lexicographically greatest element in $[q^j]$. Let us consider two other conjugates of q^j: $(p'')^2 s$ and $s(p'')^2$. We claim that $(p'')^2 s \neq p''sp''$. Otherwise, from Lemma 5, there exists a primitive word r and two positive integers $t_2 > t_1 > 0$ such that $p'' = r^{t_1}, p''s = r^{t_2}$. However, in this case, $p^j = p''s = r^{t_2}$ and $|r| < |p|$. Once more from Lemma 4, p and r cannot be both primitive. If $(p'')^2 s \prec p''sp''$, then $p''s \prec sp''$. However, if it is the case, then $p''sp'' \prec s(p'')^2$. We conclude. □

Now, we define the *Circuit Arrangement Order* $\prec_{CAO}$ on the set of small circuits in $\Gamma(w)$.

Definition 15. *Let C_1 and C_2 be two small circuits in a given Rauzy graph $\Gamma_r(w)$ for some w and r, then $C_1 \prec_{CAO} C_2$ if $e_1 \prec e_2$, where e_i is the maximal edge in C_i, $i = 1$ or 2.*

Remark 16. *Since $\prec$ is a total order on the set of words with letters in the alphabet of w, $\prec_{CAO}$ is a total order on the set of all small circuits in each Rauzy graph $\Gamma_r(w)$, $1 \leqslant r \leqslant |W|$. For any pair of small circuits C_1, C_2 in the same Rauzy graph, if $C_1 \prec_{CAO} C_2$ and if e_1, e_2 are respectively the maximal edge in C_1, C_2, then e_1 may also be an edge of C_2, but e_2 cannot be an edge of C_1. In fact, if it is the case, then e_1 cannot be the maximal edge of C_1 since $e_1 \prec e_2$, which is contradictory to the maximality of e_1.*

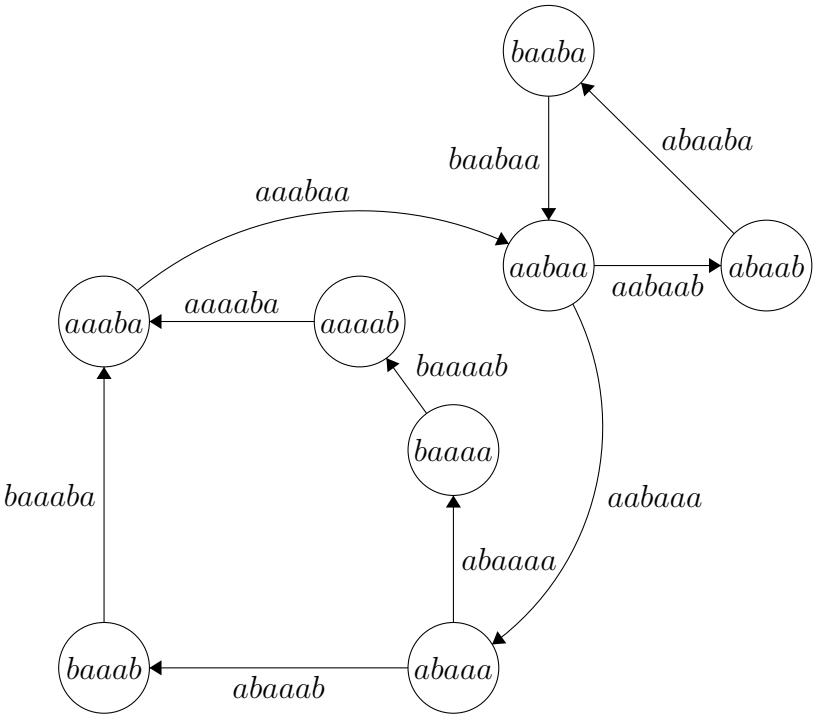
Lemma 17. *Let w be a finite word. Then for all $i, 1 \leqslant i \leqslant |w|$, the small circuits in $\Gamma_i(w)$ are independent.*

Proof. Let $C_1, C_2, \ldots, C_k$ be the small circuits in $\Gamma_i(w)$ satisfying

$$C_1 \prec_{CAO} C_2 \prec_{CAO} \cdots \prec_{CAO} C_k.$$

If there exist some real numbers $\alpha_1, \alpha_2, \ldots, \alpha_k$ such that $\sum_{t=1}^{k} \alpha_t \mu(C_t) = 0$, where $\mu(C_t)$ is the vector-cycle corresponding to C_t, then, we first prove that $\alpha_k = 0$. Let $e_{(C_k)}$ be the maximal edge of C_k. Since $C_i \prec C_k$ for all i, $1 \leqslant i < k$, from Remark 16, $e_{(C_k)}$ appears only in the circuit C_k. Thus, the component $c_{(C_k)}$ in $\mu(C_t)$ is nonzero only if $t = k$. Hence $\alpha_k = 0$. Applying this argument recursively, one has $\alpha_t = 0$ for all t, $1 \leqslant t \leqslant k$. Consequently, $C_1, C_2, \ldots, C_k$ are independent. □

Example 18. Consider the word $v = abaaabaabaaaaba$, with the order induced by $a \prec b$. The Rauzy graph $\Gamma_5(v)$ is the following:



In this graph, there are three small circuits, namely $C(aaaab, 5), C(aaab, 5)$ and $C(aab, 5)$. Further, the maximal edges of these circuits are respectively $baaaab$, $baaaba$ and $baabaa$ and appear in only one circuit, so that the three circuits are independent.

If we let $b \prec a$, then these three circuits should be denoted by $C_1 = C(baaaa, 5)$, $C_2 = C(baaa, 5)$ and $C_3 = C(baa, 5)$ and their maximal edges are respectively $e_1 = aaaaba$, $e_2 = aaabaa$ and $e_3 = aabaab$. Since $e_3 \prec e_2 \prec e_1$, $C_3 \prec_{CAO} C_2 \prec_{CAO} C_1$. Notice that although e_2 appears in both C_2 and C_1, since e_1 is only in C_1, C_1 is independent from C_2, C_3. Moreover, since e_2 is in C_2 but not in C_3, these three circuits are independent.

Lemma 19. *Let w be a finite word. For each Rauzy graph $\Gamma_i(w)$, one has*

$$sc_i(w) \leqslant |\mathrm{Fac}_{i+1}(w)| - |\mathrm{Fac}_i(w)| + 1.$$

Proof. From the facts that $\Gamma_i(w)$ is weakly connected and that the number of edges and vertices in $\Gamma_i(w)$ are respectively $\mathrm{Fac}_{i+1}(w)$ and $\mathrm{Fac}_i(w)$, the cyclomatic number of $\Gamma_i(w)$ is

$$\chi(\Gamma_i(w)) = |\mathrm{Fac}_{i+1}(w)| - |\mathrm{Fac}_i(w)| + 1,$$

and the result follows from Lemma 17 and Theorem 7. □

Lemma 20. *Let w be a finite word. The total number of small circuits in $\Gamma(w)$ is bounded by $|w| - |\mathrm{Alph}(w)|$.*

Proof. With the notation defined as above, from Lemma 19, one has:

$$\begin{aligned} sc(w) &= \sum_{i=1}^{|w|} sc_i(w) \\ &\leqslant \sum_{i=1}^{|w|} (|\mathrm{Fac}_{i+1}(w)| - |\mathrm{Fac}_i(w)| + 1) \\ &\leqslant |w| + |\mathrm{Fac}_{|w|+1}(w)| - |\mathrm{Fac}_1(w)| \\ &\leqslant |w| - |\mathrm{Alph}(w)|. \end{aligned}$$

□

4. Injection from squares to small circuits

Let w be a finite word. For any square $u^2 \in \mathrm{Fac}(w)$, there exists a primitive factor v of w and a positive integer n such that $u^2 = v^{2n}$. Thus, one can gather the squares in terms of their smallest period. For any primitive factor p of w, consider the set

$$\left\{t^{2i} | i \in \mathbb{N}^+, t \in [p], t^{2i} \in \mathrm{Fac}(w)\right\}.$$

We choose as representative an element $v \in [p]$, such that $v^{2m} \in \mathrm{Fac}(w)$, and

$$m = \max\left\{n | n \in \mathbb{N}^+, \exists t \in [p], \text{ such that } t^{2n} \in \mathrm{Fac}(w)\right\}.$$

This set will be refered as the *class of* v and denoted by $\mathrm{Class}_w(v)$, and the number m as its *index* denoted $\mathrm{Index}_w(v)$. Two classes $\mathrm{Class}_w(s)$ and $\mathrm{Class}_w(t)$ are equal only if s and t are conjugate.

Example 21. Let $w = abcabcabcabca$. The squares $(abc)^2$, $(bca)^2$, $(cab)^2$, $(abc)^4$, $(bca)^4$ are in the same class. The index of this class is 4 and this class can be denoted by $\mathrm{Class}_w(abc)$ or $\mathrm{Class}_w(bca)$. However, it cannot be denoted by $\mathrm{Class}_w(cab)$, because $(cab)^4 \notin \mathrm{Fac}(w)$.

In fact, $\mathrm{Class}_w(cab)$ does not exist. If it exists, from the definition, $\mathrm{Index}_w(cab) = 1$ and, for any conjugate v of cab and for any integer $i > 1$, $v^{2i} \notin \mathrm{Fac}(w)$. However, it is not the case, because $(abc)^4 \in \mathrm{Fac}(w)$.

Lemma 22. *Let v be a primitive factor of a finite word w such that $|v| = l$, that $\mathrm{Class}_w(v)$ is well-defined and that $\mathrm{Index}_w(v) \geqslant 1$. Then for any integer t satisfying $1 \leqslant t \leqslant |\mathrm{Class}_w(v)|$, there exists a small circuit $C(v, t+l-1)$ in the Rauzy graph $\Gamma_{t+l-1}(w)$. Hence, there exists a bijective function*

$$f_v : \mathrm{Class}_w(v) \longrightarrow \{C(v, t+l-1) | 1 \leqslant t \leqslant |\mathrm{Class}_w(v)|\},$$

such that $f_v(s) = C(v, n+l-1)$ if s is the n-th lexicographically least element in $\mathrm{Class}_w(v)$.

Proof. To establish the existence of the circuit $C(v, t+l-1)$, for any integer t in the set $\{1, \ldots, |\mathrm{Class}_w(v)|\}$, it is enough to prove that

$$[v]_{|\mathrm{Class}_{w(v)|+l}} \subset \mathrm{Fac}(w).$$

If it is the case, then

$$[v]_{t+l} \subset \text{Fac}(w), \forall t \in \{0, \dots, |\text{Class}_w(v)|\}.$$

Thus, for any $t \in \{1, \dots, |\text{Class}_w(v)|\}$, there exists a circuit in $\Gamma_{t+l-1}(w)$ such that its edge set is $[v]_{t+l}$ and its vertex set is $[v]_{t+l-1}$. Hence this circuit is $C(v, t+l-1)$. Let $c = |\text{Class}_w(v)|$, $i = \text{Index}_w(v)$ and $r = |\{u^{2i} | u^{2i} \in \text{Fac}(w), u \in [v]\}|$. It is easy to check that $r \leqslant l$ and $c = (i-1)l + r \leqslant il$. If $i > 1$, then $v^{2i} \in \text{Fac}(w)$ and $[v]_{c+l} \subset \text{Fac}(v^{2i})$. Thus, $[v]_{c+l} \subset \text{Fac}(w)$. If $i = 1$, then the elements in $\text{Class}_w(v)$ are all conjugates of v^2, thus they are all of the form $v_s(j)vv_p(j-1)$ for some j. For any $u^{\frac{c}{l}+1} \in [v]_{c+l}$, $u^{\frac{c}{l}+1} = uu_p(c)$ is a factor of the words $u_s(m)uu_p(m-1)$ for all $m \in \{c, \dots, l\}$. Hence, there are at most $c-1$ distinct words y such that y conjugates with u and that $u^{\frac{c}{l}+1} \notin \text{Fac}(y^2)$. However, there are exactly c elements in $\text{Class}_w(v)$, so there exists at least one word in $\text{Class}_w(v)$ containing $u^{\frac{c}{l}+1}$ as a factor. Thus, each element of $[v]_{c+l}$ is a factor of y^2 for some $y^2 \in \text{Class}_w(v)$. Hence $[v]_{c+l} \subset \text{Fac}(w)$.

The bijectivity of f_v follows from the fact that the cardinalities of $\text{Class}_w(v)$ and of $\{C(v, t+|v|-1) | 1 \leqslant t \leqslant |\text{Class}_w(v)|\}$ are equal. □

Lemma 23. *There exists an injective function f from the set of squares of w to the set of small circuits in $\Gamma(w)$.*

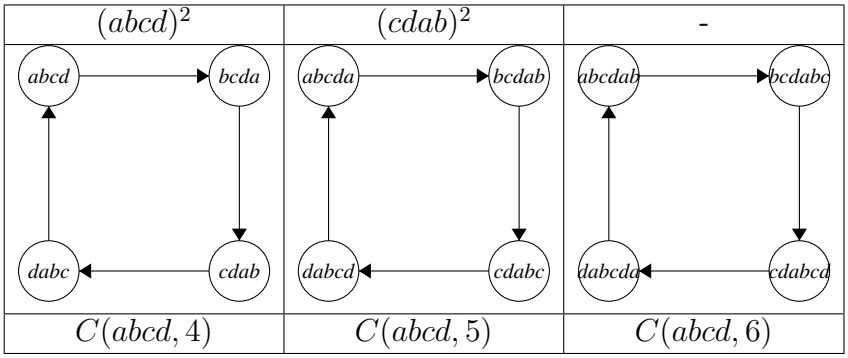
Proof. Let us consider the function f such that for each $\text{Class}(v)$,

$$f|_{\text{Class}(v)} = f_v,$$

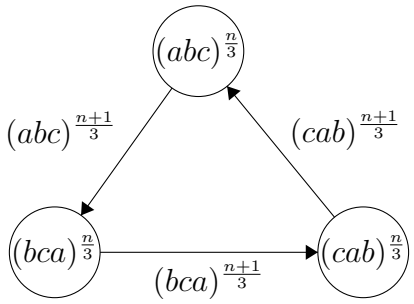
with f_v defined as in Lemma 22. We claim that this function is injective. Indeed, if $C(p,i)$ and $C(q,j)$ are the images of two squares s_1, s_2 under f such that $C(p,i) = C(q,j)$, then from Lemma 10, $p = q$ and $i = j$. Now, $p = q$ implies that s_1 and s_2 are in the same class. Further, from Lemma 22, the function f on a class is also injective, thus, $s_1 = s_2$. Hence f is injective. □

Here we illustrate the injection from squares to small circuits by considering two examples.

Example 24. Let $w = eabcdabcdecdabcdabe$, with the order $a \prec b \prec c \prec d \prec e$. There are two squares in the word w and three small circuits in the Rauzy graphs. The injection is illustrated below



Example 25. Let $w = abcabcabcabcabcabcabcabc = (abc)^8$. One can easily check that for any $3 \leqslant n \leqslant 21$, the Rauzy graph $\Gamma_w(n)$ is of the following form:



There are 10 squares in w and 19 small circuits in its Rauzy graphs. Thus, with the order $a \prec b \prec c$, there is the following injection:

$(abc)^2$	$(bca)^2$	$(cab)^2$	$(abc)^4$	$(bca)^4$
$C(abc, 3)$	$C(abc, 4)$	$C(abc, 5)$	$C(abc, 6)$	$C(abc, 7)$

$(cab)^4$	$(abc)^6$	$(bca)^6$	$(cab)^6$	$(abc)^8$
$C(abc, 8)$	$C(abc, 9)$	$C(abc, 10)$	$C(abc, 11)$	$C(abc, 12)$

This function is not a bijection since the number of squares in w is smaller than the number of small circuits in the Rauzy graphs.

Proof of Theorem 1. From Lemma 23 and Lemma 20, the number of distinct squares, $S(w)$, satisfies the inequality

$$S(w) \leqslant sc(w) \leqslant |w| - |\mathrm{Alph}(w)|. \qquad \square$$

Example 26. Consider the word $w = baababaababbbabbabbbab$ with the order $a \prec b$. Then $|w| = 22$ and there are 14 squares in w, namely

ε, aa, bb, $abab$, $baba$, $abaaba$, $bbabba$, $babbab$, $abbabb$, $babbbabb$,
$bbabbbab$, $baababaaba$, $aababaabab$, $babbbabbabbbab$.

The injection from the set of squares to the small circuits is :

a^2	b^2	$(ab)^2$	$(ba)^2$	$(aba)^2$	$(abb)^2$	$(bab)^2$
$C(a, 1)$	$C(b, 1)$	$C(ab, 2)$	$C(ab, 3)$	$C(aab, 3)$	$C(abb, 3)$	$C(abb, 4)$

$(bba)^2$	$(babb)^2$	$(bbab)^2$	$(aabab)^2$	$(baaba)^2$	$(babbbab)^2$
$C(abb, 5)$	$C(abbb, 4)$	$C(abbb, 5)$	$C(aabab, 5)$	$C(aabab, 6)$	$C(abbabbb, 7)$

5. Further discussion

The upper bound of the number of distinct squares in a word w given in Theorem 1 does not seem optimal for words such that $|w| - |\mathrm{Alph}(w)|$ is large. This is due to two facts: first, the injection established in Section 4 is not a bijection when there exists some factors $u^k \in \mathrm{Fac}(w)$ such that $k \geqslant 3$; second, there could be other circuits or cycles which are independent from the small circuits. For example, if we consider the word $w = (abc)^8$ given in Example 25, there exists a circuit in $\Gamma_1(w)$ such that its vertices are $\{a, b, c\}$ and its edges are $\{ab, bc, ca\}$. Clearly

this circuit is not small but it is independent from all small circuits in $\Gamma(w)$. In fact, the maximum number of distinct squares in a binary string of length n has been computed for small n's and it is listed as A248958 in OEIS [OEI23]. Motivated by this computation, we suggest a stronger upper bound for the number of distinct squares.

Conjecture 27. *For any finite word w of length k, the number $S(w)$ of its distinct square factors satisfies*

$$S(w) \leqslant \left\lceil k - \sqrt{k} - \log_2\left(\sqrt{k}\right) \right\rceil,$$

where $\lceil x \rceil$ denotes the smallest integer larger than or equal to x.

Acknowledgements

The authors would like to express their gratitude towards the referees for careful reading and suggestions to improve the paper, as well as suggestions for other references.

References

[AA07] Boris Adamczewski and Jean-Paul Allouche. Reversals and palindromes in continued fractions. *Theoretical Computer Science*, 380(3):220–237, 2007. doi:10.1016/j.tcs.2007.03.017.

[ABCD03] Jean-Paul Allouche, Michael Baake, Julien Cassaigne, and David Damanik. Palindrome complexity. *Theoretical Computer Science*, 292(1):9–31, 2003. doi:10.1016/s0304-3975(01)00212-2.

[Ber82] C. Berge. *The Theory of Graphs and Its Applications*. Greenwood Press, 1982.

[BGL13] Alexandre Blondin Massé, Ariane Garon, and Sébastien Labbé. Combinatorial properties of double square tiles. *Theoretical Computer Science*, 502:98–117, 2013. doi:10.1016/j.tcs.2012.10.040.

[dB46] N. G. de Bruijn. A combinatorial problem. *Koninklijke Nederlandse Akademie v. Wetenschappen*, 49:758–764, 1946.

[DF14] Antoine Deza and Frantisek Franek. A d-step approach to the maximum number of distinct squares and run in strings. *Discrete Applied Mathematics*, 163:268–274, 2014. doi:10.1016/j.dam.2013.10.021.

[DFJ11] Antoine Deza, Frantisek Franek, and Mei Jiang. A d-step approach for distinct squares in strings. In Raffaele Giancarlo and Giovanni Manzini, editors, *Proceedings of Combinatorial Pattern Matching*, volume 6661 of *Lecture Notes in Computer Science*, pages 77–89. Springer Berlin / Heidelberg, 2011. doi:10.1007/978-3-642-21458-5_9.

[DFT15] Antoine Deza, Frantisek Franek, and Adrien Thierry. How many double squares can a string contain? *Discrete Applied Mathematics*, 180:52–69, 2015. doi:10.1016/j.dam.2014.08.016.

[FS98] Aviezri S. Fraenkel and Jamie Simpson. How many squares can a string contain? *J. Comb. Theory, Ser. A*, 82(1):112–120, 1998. doi:10.1006/jcta.1997.2843.

[FW65] N. J. Fine and H. S. Wilf. Uniqueness theorems for periodic functions. *Proceedings of the American Mathematical Society*, 16(1):109–114, 1965. doi:10.2307/2034009.

[Ili07] Lucian Ilie. A note on the number of squares in a word. *Theor. Comput. Sci.*, 380(3):373–376, 2007. doi:10.1016/j.tcs.2007.03.025.

[Lam13] N. H. Lam. On the number of squares in a string. Technical Report 2, AdvOL-Report, McMaster University, 2013.

[Lot05] M. Lothaire. *Applied Combinatorics on Words*. Cambridge University Press, Cambridge, 2005.

[LS62] R. C. Lyndon and M. P. Schützenberger. The equation $a^M = b^N c^P$ in a free group. *Michigan Mathematical Journal*, 9(4):289 – 298, 1962. doi:10.1307/mmj/1028998766.

[OEI23] OEIS Foundation Inc. The On-Line Encyclopedia of Integer Sequences, 2023. Published electronically at http://oeis.org.

[Rau83] G. Rauzy. Suites à termes dans un alphabet fini. *Seminaire de Théorie des Nombres de Bordeaux*, 12:1–16, 1983. URL: http://eudml.org/doc/182163.

[Thi20] A. Thierry. A proof that a word of length n has less than 1.5n distinct squares. 2020. arXiv:2001.02996.