View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Dagstuhl Research Online Publication Server

On Maximal Repeats in Compressed Strings

Julian Pape-Lange
Technische Universität Chemnitz, Straße der Nationen 62, 09111 Chemnitz, Germany
julian.pape-lange@informatik.tu-chemnitz.de

Abstract

This paper presents and proves a new non-trivial upper bound on the number of maximal repeats of compressed strings. Using Theorem 1 of Raffinot's article "On Maximal Repeats in Strings", this upper bound can be directly translated into an upper bound on the number of nodes in the Compacted Directed Acyclic Word Graphs of compressed strings.

More formally, this paper proves that the number of maximal repeats in a string with z (self-referential) LZ77-factors and without q-th powers is at most $3q(z+1)^3 - 2$. Also, this paper proves that for $2000 \leq z \leq q$ this upper bound is tight up to a constant factor.

2012 ACM Subject Classification Mathematics of computing → Combinatorics on words; Mathematics of computing → Combinatoric problems

Keywords and phrases Maximal repeats, Combinatorics on compressed strings, LZ77, Compact suffix automata, CDAWGs

Digital Object Identifier 10.4230/LIPIcs.CPM.2019.18

Acknowledgements I thank Professor Laurent Bartholdi for leading me to this research topic as well as for his helpful and inspiring pieces of advice.

1 Introduction

A repeat of a string S is a substring of S which occurs at least twice in S. A repeat P of S is a maximal repeat, if every string which properly contains P occurs less often in S than P itself. Usually there are much less maximal repeats than repeats. Nevertheless the set of maximal repeats still contains all of the information about the repeats. These repeats have, as shown by Gusfield in [9], many applications in computational biology. A good overview of the importance of repeats in computational biology together with a deeper analysis of local repeats is also given by Nicolas et al. in [10] on ResearchGate.

Maximal repeats are also closely linked to string compression and succinct data structures: Furuya et al. show in their recent arXiv-article [8] that there is a connection between maximal repeats and the grammar compression algorithm RePair and they use this connection to create an improved version of this algorithm. Raffinot proves in [12] that there is a natural one-to-one correspondence between the maximal repeats of a string and the number of internal nodes in its Compacted Directed Acyclic Word Graph (CDAWG).

The CDAWG of a string is a useful data structure which was introduced by Blumer et al. in [2] and has most advantages of suffix trees and acyclic directed word graphs while usually being much smaller than each of them. The CDAWG is therefore a powerful tool for string processing.

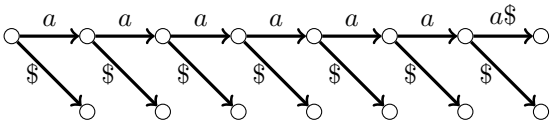

Figure 1 The suffix tree of a^7 $(= aaaaaaa)$.

© Julian Pape-Lange;
licensed under Creative Commons License CC-BY
30th Annual Symposium on Combinatorial Pattern Matching (CPM 2019).
Editors: Nadia Pisanti and Solon P. Pissis; Article No. 18; pp. 18:1–18:13
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

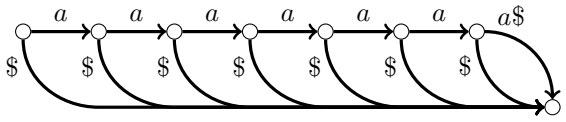

Figure 2 The CDAWG of a^7.

One might hope that well-compressible strings have highly structured suffix trees and thereby small CDAWGs. This, however, is unfortunately not the case. Even the arguably best compressible string, a^{q-1}, which does have the simple looking suffix tree shown in Figure 1, also has the CDAWG shown in Figure 2 with $q-2$ internal nodes. This shows that there is no non-trivial upper bound on either the number of nodes of a CDAWG of a string S or the number of maximal repeats of S which is only dependent on the compressed size of S. This fact may also explain the apparent lack of research regarding the number of nodes in CDAWGs of general compressible strings.

There are, however, some non-trivial bounds for the number of nodes in CDAWGs which take the structure of the underlying strings into account. For example Blumer et al. suggest in [2] that the number of nodes in a CDAWG of an English lower-case string S is between 0.26 times the length of S and 0.29 times the length of S. Blumer et al. prove in [3] formulas for the average size of the CDAWG of a random string. Stronger results have been found by Radoszewski and Rytter who prove in [11] that the number of nodes of the CDAWG of Thue-Morse words is linear in the compressed size of the word and thereby logarithmic in the size of the word itself. A similar result is shown by Epifanio et al. in [6] for Sturmian words.

Belazzougui et al. prove in [1] that the number of edges in the CDAWG of the string is bounded from below by the number of self-referential LZ77-factors. Therefore a string S over the alphabet Σ with z LZ77-factors has at least $\frac{z}{|\Sigma|}-1$ maximal repeats. This lower bound is met, for example, by a string in which every character occurs only once. While this string is not compressible, it does not have maximal repeats.

While Raffinot was motivated by the possibility of translating the better-known results for CDAWGs to maximal repeats, this paper's motivation was the other way round. The main goal was to find a new, more general upper bound for the number of nodes in the CDAWGs of compressed strings and it turned out to be very useful that Raffinot's result can be applied the other way round too.

The number z of LZ77-factors proved itself to be a very useful indicator of the complexity of strings in the past. For example Charikar et al. proved in [4] that even the minimal number of non-self-referential LZ77-factors is a lower bound for the smallest grammar compression. The self-referential version of LZ77 was used for example by Tanimura in [13] in order to show that the size of the t-truncated suffix tree is bounded by zt. Additionally, since high powers lead to CDAWGs with a high number of nodes, the additional structure of the string is measured by the highest power $q-1$ in the string.

Using these two variables, this paper gives an upper bound for the number of maximal repeats and the number of nodes in the CDAWG which is proven in section 3:

▶ **Theorem 1.** *Let S be a string. Let z be the number of (self-referential) LZ77-factors in an LZ77-decomposition of S. Let q be a number such that S does not contain q-th powers. Then the number of maximal repeats in S is bounded from above by $3q(z+1)^3-2$. Also, the Compacted Directed Acyclic Word Graph (CDAWG) of S has at most $3q(z+1)^3$ nodes.*

Additionally this paper shows:

▶ **Theorem 2.** *For $2000 \leq z \leq q$ there is a string S without q-th powers which can be expressed by z (self-referential) LZ77-factors and which has at least $\frac{1}{500}qz^3$ maximal repeats.*

This result, which is proven in section 4, shows that for $2000 \leq z \leq q$ the upper bound given by Theorem 1 is tight up to a constant factor.

2 Definitions

Let Σ be an *alphabet*. A *string* with *length* denoted by $|S|$ is the concatenation of *characters* $S[0]S[1]\cdots S[|S|-1]$ of Σ. For the sake of convenience we also define $S[-1] = \$$ and $S[|S|] = \$$ with $\$ \notin \Sigma$. The *substring* $S[i..j]$ with $0 \leq i \leq j \leq |S|-1$ is the concatenation $S[i]S[i+1]\cdots S[j]$. For $i > j$ the substring $S[i..j]$ is defined to be the empty string with length 0. A *prefix* is a substring of the form $S[0..j]$ and a *suffix* is a substring of the form $S[i..|S|-1]$.

A *maximal pair* of S is a triple $(n, m, l) \in \mathbb{N}^3$ with $l \geq 1$ such that $S[n..n+l-1]$ is equal to $S[m..m+l-1]$ and this property can not be extended to any side. More formally:

- $\forall i \in \mathbb{N}$ with $0 \leq i < l : S[n+i] = S[m+i]$ but
- $S[n-1] \neq S[m-1]$ and
- $S[n+l] \neq S[m+l]$.

Since for a maximal pair (n, m, l) the inequality $S[n-1] \neq S[m-1]$ holds, the indices n and m can not be equal. Furthermore, only $S[n..n+l-1]$ and $S[m..m+l-1]$ are required to be in S. The characters $S[n-1]$, $S[m-1]$, $S[n+l]$, $S[m+l]$ may be outside of S. This implies that $S[n..n+l]$ and $S[m..m+l]$ are in $S\$$.

The *distance* d of a maximal pair (n, m, l) is the distance $d = m - n$ of the two starting indices.

A *maximal repeat* of a string S is a substring $S[n..n+l-1]$ such that there is a maximal pair (n, m, l) for some indices n, m.

For example, in the string *banana*, the substring *na* is not a maximal repeat, because every occurrence of *na* is preceded by *a*. The substring *ana*, however, is a maximal repeat with maximal pair given by $(1, 3, 3)$. The distance of this maximal pair is 2.

A (self-referential) *LZ77-decomposition* of a string S is a factorization $S = F_1F_2\ldots F_z$ in *LZ77-factors* $F_1, F_2, \ldots, F_z$ such that for all $i \in 1, 2, \ldots, z$

- the factor F_i is a single character or
- the substring F_i occurs twice in $F_1F_2\ldots F_i$. (i.e. there is an occurrence of F_i in $F_1F_2\ldots F_i$ which does not use the last character of $F_1F_2\ldots F_i$)

In this paper, all LZ77-decompositions are allowed to be self-referential. Therefore we will only use the term LZ77-decomposition.

Normally the LZ77-definition requires the number of LZ77-factors of a string to be minimized. Since all theorems of this paper also hold for non-minimized LZ77-decompositions, this minimization is not required in this paper.

For example, see the following strings on the left-hand side with possible corresponding LZ77-factors, separated by "$\cdot$", on the right-hand side:

- $01001010 = 0 \cdot 1 \cdot 0 \cdot 010 \cdot 10$,
- $banana = b \cdot a \cdot n \cdot ana$,
- $aaaa = a \cdot aaa$ and
- $aaaa = a \cdot a \cdot a \cdot a$ (not minimal).

A *period* of a string S is a number Δ such that all characters in S with distance Δ are equal. If the minimal period $\Delta_{\min}$ of a non-empty string S is at most $\frac{|S|}{2}$, the string S is a *fractional power* with *exponent* $\frac{|S|}{\Delta_{\min}}$.

Fractional powers are also called repetitions in the literature. However, in order to keep them apart from the maximal repeats, the name fractional power will be used.

3 Upper Bound

The main goal of this section is to prove that the number of maximal repeats of a string S that can be written with z LZ77-factors and that does not contain a q-th power is bounded from above by $3q(z+1)^3 - 2$.

While it is easier to count the number of maximal pairs than to count the number of maximal repeats directly, there might be many maximal pairs for a single maximal repeat. Therefore, it is necessary to choose a subset of the maximal pairs which presents every maximal repeat at least once and which does not contain too many elements.

The following two lemmata will lead to a suitable subset of the maximal pairs, by showing that it is sufficient to count the maximal pairs (n, m, l) in which n is smaller than m and n as well as m are close to the boundary between two LZ77-factors.

▶ **Lemma 3.** *The triple (n, m, l) is a maximal pair if and only if (m, n, l) is a maximal pair.*

Proof. This lemma follows directly from the symmetry of the definition of maximal pairs. ◀

▶ **Lemma 4.** *Let S be a string. Let $F_1F_2 \dots F_zF_{z+1} = S\$$ be an LZ77-decomposition of $S\$$ and $s_1, s_2, \dots, s_z, s_{z+1}$ be the starting indices of the LZ77-factors in $S\$$. Let (n, m, l) be a maximal pair in S. Then there is a maximal pair (n', m', l) such that the equation $S[n..n+l-1] = S[n'..n'+l-1]$ holds and the intervals $[n', n'+l]$ and $[m', m'+l]$ contain starting indices s_j and s_k respectively.*

Proof. Let n' and m' the minimal indices such that $S[n-1..n+l] = S[n'-1..n'+l]$ and $S[m-1..m+l] = S[m'-1..m'+l]$. By construction (n', m', l) is a maximal pair and $S[n..n+l-1] = S[n'..n'+l-1]$ holds.

Assume the interval $[n'-1, n'+l]$ is inside an interval $[s_i, s_i+|F_i|-1]$ and thereby inside the LZ77-factor F_i.

Since the interval contains more than one character, every substring of F_i has an earlier occurrence. This contradicts the minimality of n'.

Therefore the last index in the interval $[n'-1, n'+l]$ lies inside another LZ77-factor than the first index in this interval. This implies the interval $[n', n'+l]$ contains some starting index s_j. Similarly, the interval $[m', m'+l]$ contains some starting index s_k. ◀

The next two lemmata will show some properties of maximal pairs with overlap. These properties will be important in the proof of the upper bound for the subset of maximal pairs.

▶ **Lemma 5.** *Let S be a string. Let further (n_a, m_a, l_a) and (n_b, m_b, l_b) be different maximal pairs in S such that there is an index c with $c \in [n_a, n_a+l_a]$ and $c \in [n_b, n_b+l_b]$. Then the distances $d_a = m_a - n_a$ and $d_b = m_b - n_b$ are unequal.*

Proof. By contradiction:

Assume The equation $d_a = d_b$ holds:

This implies $n_a - n_b = m_a - m_b$

Case 1: $n_a = n_b$ (see for example Figure 3):

Using $n_a = n_b$ and thereby $m_a = m_b$, it follows that $l_a \neq l_b$ holds. This, however implies

$$0 < \min(l_a, l_b) < \max(l_a, l_b) \tag{1}$$

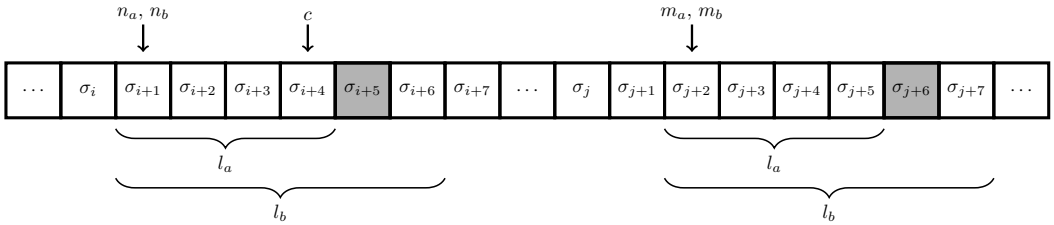

Figure 3 Case 1 of Lemma 5. The characters σ_{i+5} and σ_{j+6} have to be unequal because of their position just outside (n_a, m_a, l_a) and have to be equal because of their position in (n_b, m_b, l_b).

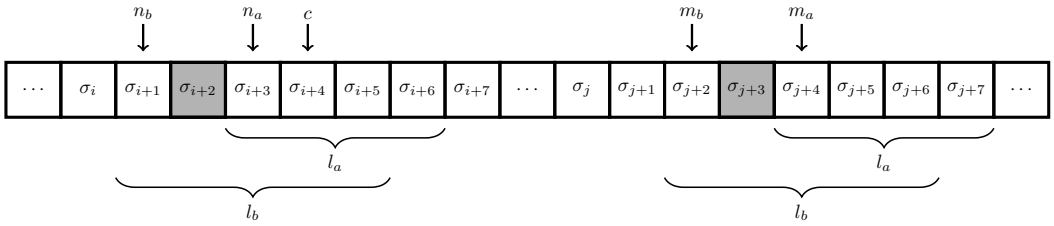

Figure 4 Case 2 of Lemma 5. The characters σ_{i+2} and σ_{j+3} have to be unequal because of their position just outside (n_a, m_a, l_a) and have to be equal because of their position in (n_b, m_b, l_b).

and thereby

$$S[n_a + \min(l_a, l_b)] \overset{1}{=} S[m_a + \min(l_a, l_b)].$$

Hence, either (n_a, m_a, l_a) or (n_b, m_b, l_b) is not a maximal pair.
Therefore this case is not possible.

Case 2: $n_a \neq n_b$ (see for example Figure 4):
Without loss of generality $n_a > n_b$ holds. Since $n_a \leq c$ and $c \leq n_b + l_b$ hold, the inequality

$$0 \leq n_a - n_b - 1 < l_b \tag{2}$$

follows and using $n_a - n_b = m_a - m_b$ we get

$$\begin{aligned} S[n_a - 1] = S[n_b + (n_a - n_b - 1)] &\overset{2}{=} S[m_b + (n_a - n_b - 1)] \\ &= S[m_b + (m_a - m_b - 1)] = S[m_a - 1] \end{aligned}$$

Hence, (n_a, m_a, l_a) is not a maximal pair.
Therefore this case is not possible.

Since all cases contradict the assumption, the distances d_a and d_b are unequal. ◀

▶ **Lemma 6.** *Let S be a string. Let further (n_a, m_a, l_a) and (n_b, m_b, l_b) be maximal pairs in S with distances $d_a \neq d_b$. Define the difference of the distances $\Delta_d = d_a - d_b$. Then $S[\max(n_a, n_b)..\min(n_a + l_a, n_b + l_b) - 1]$ is $|\Delta_d|$-periodic.*

Proof. Without loss of generality $n_a \geq n_b$ holds. Then $\max(n_a, n_b) = n_a$ holds and the string $S[\max(n_a, n_b)..\min(n_a + l_a, n_b + l_b) - 1]$ has length $\min(l_a, n_b - n_a + l_b)$.

Let x be a natural number such that $0 \leq x < x + |\Delta_d| < \min(l_a, n_b - n_a + l_b)$ holds.

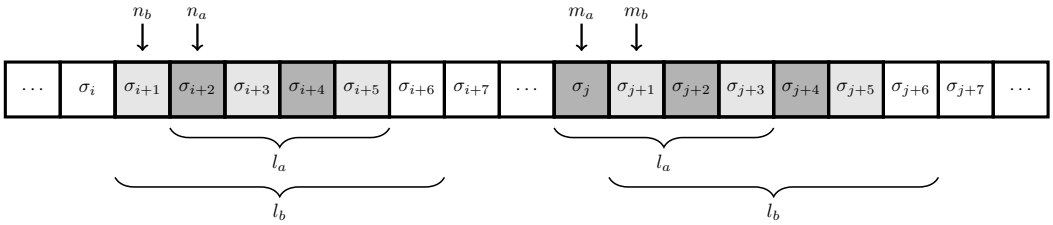

Figure 5 Case 1 of Lemma 6. The characters σ_{i+2} and σ_{j+2} have to be equal because of their position in (n_b, m_b, l_b) and the characters σ_{j+2} and σ_{i+4} have to be equal because of their position in (n_a, m_a, l_a).

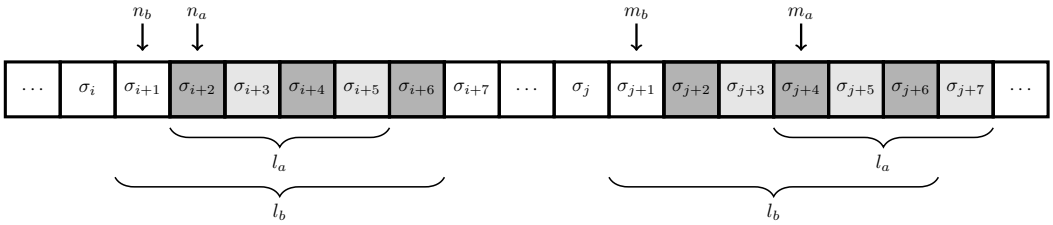

Figure 6 Case 2 of Lemma 6. The characters σ_{i+2} and σ_{j+4} have to be equal because of their position in (n_a, m_a, l_a) and the characters σ_{j+4} and σ_{i+4} have to be equal because of their position in (n_b, m_b, l_b).

Case 1: $\Delta_d < 0$ (see for example Figure 5):

In this case

$$0 \leq n_a - n_b \leq x + (n_a - n_b)$$
$$x + (n_a - n_b) < n_b - n_a + l_b + (n_a - n_b) = l_b \tag{3}$$

and

$$0 < x + |\Delta_d| < l_a \tag{4}$$

hold. Therefore

$$\begin{aligned} S[n_a + x] &= S[n_b + x + (n_a - n_b)] \\ &\overset{3}{=} S[m_b + x + (n_a - n_b)] \\ &= S[m_a + x - (m_a - m_b) + (n_a - n_b)] \\ &= S[m_a + x - \Delta_d] \\ &= S[m_a + x + |\Delta_d|] \\ &\overset{4}{=} S[n_a + x + |\Delta_d|] \end{aligned}$$

holds.

Case 2: $\Delta_d > 0$ (see for example Figure 6):

In this case

$$0 \leq x < l_a \tag{5}$$

and

$$0 < x + |\Delta_d| = x + \Delta_d = x + (m_a - m_b) - (n_a - n_b) \leq x + (m_a - m_b)$$
$$\begin{aligned} x + (m_a - m_b) &= x + \Delta_d + (n_a - n_b) \\ &= x + |\Delta_d| + (n_a - n_b) < n_b - n_a + l_b + (n_a - n_b) = l_b \end{aligned} \tag{6}$$

hold. Therefore

$$\begin{aligned} S[n_a + x] &\overset{5}{=} S[m_a + x] \\ &= S[m_b + x + (m_a - m_b)] \\ &\overset{6}{=} S[n_b + x + (m_a - m_b)] \\ &= S[n_a + x + (m_a - m_b) - (n_a - n_b)] \\ &= S[n_a + x + \Delta_d] \\ &= S[n_a + x + |\Delta_d|] \end{aligned}$$

holds.

Therefore for all numbers x with $0 \leq x < x + |\Delta_d| < \min(l_a, n_b - n_a + l_b)$ the equation $S[n_a+x] = S[n_a+x+|\Delta_d|]$ holds. Therefore the string $S[\max(n_a, n_b)..\min(n_a+l_a, n_b+l_b)-1]$ is $|\Delta_d|$-periodic. ◀

To use the periodicities we will utilize the following lemma. The simplification used here was presented in the book of Crochemore and Rytter in [5]. The original Lemma comes from the article [7] of Fine and Wilf.

▶ **Lemma 7** (Weak Periodicity Lemma). *Let P be a string with periods Δ_1 and Δ_2 such that $\Delta_1 + \Delta_2 \leq |P|$. Then $\gcd(\Delta_1, \Delta_2)$ is a period of P.*

With all this preparation it is now possible to count maximal pairs around given indices:

▶ **Theorem 8.** *Let S be a string. Let $F_1F_2 \dots F_zF_{z+1} = S\$$ be an LZ77-decomposition of $S\$$. Let $s_1, s_2, \dots, s_z, s_{z+1}$ be the starting indices of the LZ77-factors in $S\$$. Let $q \in \mathbb{N}_{\geq 2}$ and $i, j \in \{1, 2, \dots, z, z+1\}$ be natural numbers.*
Then the number of different maximal pairs (n_k, m_k, l_k) such that for all k

- *the substring $S[n_k..s_i - 1]$ is not a fractional power with exponent greater than or equal to q,*
- *the substring $S[s_i..n_k + l_k - 1]$ is not a fractional power with exponent greater than or equal to q,*
- *the starting index s_i is contained in the interval $[n_k, n_k + l_k]$,*
- *the starting index s_{i+1} is not contained in the interval $[n_k, n_k + l_k]$ and*
- *the starting index s_j is contained in the interval $[m_k, m_k + l_k]$*

is bounded from above by $18q \cdot \lceil \log_q(|F_1F_2 \dots F_i|) \rceil$

Proof. By contradiction:

Assume there are at least $(18q \cdot \lceil \log_q(|F_1F_2 \dots F_i|) \rceil) + 1$ different maximal pairs with the restrictions given by the prerequisites:

We will now use the pigeonhole principle until we get two pairs of maximal pairs which have a huge overlap and similar distances.

For each of these maximal pairs (n_k, m_k, l_k) at least one of the following options hold:

- At least half of the interval $[n_k, n_k + l_k - 1]$ lies before s_i (i.e. $n_k + \frac{l_k}{2} \leq s_i$), or
- At least half of the interval $[n_k, n_k + l_k - 1]$ lies after $s_i - 1$ (i.e. $n_k + \frac{l_k}{2} \geq s_i$).

Since there are two options and at least $(18q \cdot \lceil \log_q(|F_1F_2 \dots F_i|) \rceil) + 1$ maximal pairs at least one of these options hold for

$$\left\lceil \frac{(18q \cdot \lceil \log_q(|F_1F_2 \dots F_i|) \rceil) + 1}{2} \right\rceil = (9q \cdot \lceil \log_q(|F_1F_2 \dots F_i|) \rceil) + 1$$

maximal pairs. By symmetry we can assume without loss of generality that there are $(9q \cdot \lceil \log_q(|F_1F_2 \dots F_i|) \rceil) + 1$ of the given maximal pairs satisfying $n_k + \frac{l_k}{2} \leq s_i$.

Since all of these $(9q \cdot \lceil \log_q(|F_1F_2 \dots F_i|) \rceil) + 1$ maximal pairs (n_k, m_k, l_k) satisfy both $s_i \in [n_k, n_k + l_k]$ and $s_{i+1} \notin [n_k, n_k + l_k]$, the inequality $l_k \leq |F_1F_2 \dots F_i|$ holds.

Taking the logarithm yields

$$0 = log_q(1) \leq log_q(l_k) \leq log_q(|F_1F_2 \dots F_i|) \leq \lceil \log_q(|F_1F_2 \dots F_i|) \rceil.$$

Since every $log_q(l_k)$ lies in at least one of the $\lceil \log_q(|F_1F_2 \dots F_i|) \rceil$ intervals $[h, h+1]$ with $0 \leq h < \lceil \log_q(|F_1F_2 \dots F_i|) \rceil$, the pigeonhole principle yields that there has to be a natural number L' such that

$$\left\lceil \frac{(9q\lceil \log_q(|F_1F_2 \dots F_i|) \rceil) + 1}{\lceil \log_q(|F_1F_2 \dots F_i|) \rceil} \right\rceil \geq 9q + 1$$

of these maximal pairs have length $L' \leq log_q(l_k) \leq 1 + L'$.

For $L = q^{L'}$ this gives a natural number L such that $L \leq l_k \leq qL$ holds for these $9q + 1$ maximal pairs.

Therefore there is a real number θ such that

- for at least $3q + 1$ of these $9q + 1$ maximal pairs $L \leq l_k \leq \theta L$ holds and
- for at least $6q + 1$ of these $9q + 1$ maximal pairs $\theta L \leq l_k \leq qL$ holds.

With the given restrictions $s_i \in [n_k, n_k + l_k]$ and $s_j \in [m_k, m_k + l_k]$ from the main assumption as well as $n_k + \frac{l_k}{2} \leq s_i$ from the application of the pigeonhole principle it follows that $n_k + \frac{l_k}{2} \leq s_i \leq n_k + l_k$ and $m_k \leq s_j \leq m_k + l_k$ hold. Therefore

$$s_j - s_i - \frac{l_k}{2} \leq (m_k + l_k) - \left(n_k + \frac{l_k}{2}\right) - \frac{l_k}{2} = m_k - n_k = d_k \text{ and}$$

$$d_k = m_k - n_k = m_k - (n_k + l_k) + l_k \leq s_j - s_i + l_k$$

hold and d_k lies in the interval $[s_j - s_i - \frac{l_k}{2}, s_j - s_i + l_k]$.

Of the $6q + 1$ maximal pairs (n_k, m_k, l_k) with $\theta L \leq l_k \leq qL$, each d_k is in at least one of the $6q$ intervals $[s_j - s_i - \frac{qL}{2} + h \cdot \frac{1}{4}L, s_j - s_i - \frac{qL}{2} + (h+1)\frac{1}{4}L]$ with $0 \leq h < 6q$. Therefore, the pigeonhole principle yields that at least

$$\left\lceil (6q+1)\frac{\frac{1}{4}L}{\frac{3}{2}qL} \right\rceil = \left\lceil (6q+1)\frac{1}{6q} \right\rceil = 2$$

of these maximal pairs have distances d_a, d_b with $|d_a - d_b| \leq \frac{1}{4}L$. Using Lemma 5 and Lemma 6 as well as $n_k + \frac{\theta L}{2} \leq n_k + \frac{l_k}{2} \leq s_i$ and $s_i \leq n_k + l_k$ for all these maximal pairs, we obtain that there is a maximal pair $(n_\alpha, m_\alpha, l_\alpha)$ such that $n_\alpha \leq s_i - \frac{\theta}{2}L$ and such that $S[n_\alpha..s_i - 1]$ has a period of $0 < \Delta_\alpha \leq \frac{1}{4}L$.

Similarly it can be shown that of the $3q + 1$ maximal pairs (n_k, m_k, l_k) with $L \leq l_k \leq \theta L$, there is a maximal pair $(n_\beta, m_\beta, l_\beta)$ such that $n_\beta \leq s_i - \frac{1}{2}L$ and such that $S[n_\beta..s_i - 1]$ has a period of $0 < \Delta_\beta \leq \frac{\theta}{2q}L$.

Since $S[n_\alpha..s_i - 1]$ is not a fractional power with exponent greater than or equal to q, we obtain $\frac{s_i - n_\alpha}{\Delta_\alpha} < q$. With $\frac{\theta L}{2} \leq \frac{l_\alpha}{2} \leq s_i - n_\alpha$ and $\Delta_\alpha \leq \frac{1}{4}L$ it follows that $\theta < \frac{q}{2}$ and hence $\Delta_\beta \leq \frac{1}{4}L$ hold.

Since $S[\max(n_\alpha, n_\beta)..s_i - 1]$ has length of at least $\frac{1}{2}L$ and is Δ_α-periodic as well as Δ_β-periodic with $\Delta_\alpha + \Delta_\beta \leq \frac{1}{2}L$ the periodicity lemma is applicable and shows that $S[\max(n_\alpha, n_\beta)..s_i - 1]$ is $\gcd(\Delta_\alpha, \Delta_\beta)$-periodic. This implies that $S[s_i - \Delta_\alpha..s_i - 1]$ is $\gcd(\Delta_\alpha, \Delta_\beta)$-periodic. Since $S[n_\alpha..s_i - 1]$ is Δ_α-periodic and at least one substring with length Δ_α is $\gcd(\Delta_\alpha, \Delta_\beta)$-periodic, even $S[n_\alpha..s_i - 1]$ is $\gcd(\Delta_\alpha, \Delta_\beta)$-periodic.

However with $\gcd(\Delta_\alpha, \Delta_\beta) \leq \Delta_\beta \leq \frac{\theta}{2q}L$ this implies that the substring $S[n_\alpha..s_i - 1]$ with at least $\frac{\theta}{2}L$ characters has a period with length of at most $\frac{\theta}{2q}L$. Therefore $S[n_\alpha..s_i - 1]$ is a fractional power with exponent greater than or equal to q.

This however contradicts the assumption and thereby proves the theorem. ◀

Now it is time to prove Theorem 1 which was stated in the introduction:

▶ **Theorem 1.** *Let S be a string. Let z be the number of (self-referential) LZ77-factors in an LZ77-decomposition of S. Let q be a number such that S does not contain q-th powers. Then the number of maximal repeats in S is bounded from above by $3q(z+1)^3 - 2$. Also, the Compacted Directed Acyclic Word Graph (CDAWG) of S has at most $3q(z+1)^3$ nodes.*

Proof. Lemma 3 shows that it is sufficient to count maximal pairs (n_k, m_k, l_k) with $n_k < m_k$. Lemma 4 shows that we can additionally require $s_i \in [n_k, n_k + l_k]$ and $s_j \in [m_k, m_k + l_k]$ for some starting indices s_i and s_j of the $z+1$ LZ77-factors of the string $S\$$.

Since the first LZ77-factor is always a single character, the equation $|F_1| = 1 = q^0$ holds. Since S does not contain a q-th power, every LZ77-factor can at most multiply the length of the string by the factor q. Therefore $|F_1F_2 \dots F_i| \leq q|F_1F_2 \dots F_{i-1}|$ holds. Induction therefore yields $|F_1F_2 \dots F_i| \leq q^{i-1}$. This implies $\lceil \log_q(|F_1F_2 \dots F_i|) \rceil \leq i - 1$

Since \$ does not occur in S, the last LZ77-factor of $S\$$ consists of only the character $\$ = S[|S|]$. Since $n_k < m_k \leq |S| - l_k$ holds, the inequality $n_k + l_k < |S|$ holds as well. This implies that s_{z+1} is not contained in $[n_k, n_k + l_k]$.

Using Theorem 8 and summing up over all pairs (s_i, s_j) with $s_i \leq s_j$ and $s_i \leq s_z$ yield that there are at most

$$\begin{aligned}
\sum_{i=1}^{z}\sum_{j=i}^{z+1}\left(18q \cdot \lceil \log_q(|F_1F_2 \dots F_i|) \rceil\right) &\leq 18q\sum_{i=1}^{z}\sum_{j=i}^{z+1}(i-1) \\
&= 18q\sum_{i=1}^{z}(i-1)(z+2-i) \\
&= 18q\sum_{i=1}^{z}(-i^2 + i(z+3) - (z+2)) \\
&= 3q(z^3 + 3z^2 - 4z) \\
&\leq 3q(z+1)^3 - 2
\end{aligned}$$

maximal repeats in S.

Raffinot shows in Theorem 1 of [12] that the maximal repeats of a string S are exactly the representatives of the internal states of the CDAWG of S. This implies that the CDAWG of S has at most $3q(z+1)^3$ states. ◀

4 Tightness

The goal of this section is to prove that for every q, z with $2000 \leq z \leq q$ there are strings without q-th powers which can be described with z LZ77-factors and which have at least $\frac{1}{500}qz^3$ maximal repeats. This also proves that the upper bound given in the last section can not be improved by more than a constant factor.

The proof of Theorem 8 suggests that high powers are necessary in order to have many maximal repeats. We therefore create a string $V_{v,q,q}$ consisting of nested $2q$-th powers first and then build a bigger string consisting of $V_{v,q,q}$ and some shortened copies of $V_{v,q,q}$.

We therefore define for natural numbers v, d, q and c with $v \geq 1$ and $d \leq q$:

$$V_{0,*,*} := \sigma_0,$$
$$V_{v,c,q} := (V_{v-1,q,q})^c \sigma_v (V_{v-1,q,q})^c,$$
$$L_{v,c,q} := (V_{v-1,q,q})^c \sigma_v (V_{v-1,q,q})^c,$$
$$R_{v,c,q} := (V_{v-1,q,q})^q \sigma_v (V_{v-1,q,q})^c,$$
$$C_{v,c,q} := L_{1,c,q} L_{2,c,q} \dots L_{v-1,c,q} V_{v,c,q} R_{v-1,c,q} \dots R_{2,c,q} R_{1,c,q} \text{ and}$$
$$S_{v,d,q} := V_{v,q,q} \left(\prod_{i=1}^{d} \$ C_{v,q-i,q} \right).$$

In order to find the highest power and the number of LZ77-factors, it is first necessary to show, that the $C_{v,q-i,q}$ are indeed proper substrings of $V_{v,q,q}$.

▶ **Lemma 9.** *For $c \leq q-1$ the string $L_{1,c,q} L_{2,c,q} \dots L_{w,c,q}$ is a proper suffix of $V_{w,q,q}$ and the string $R_{w,c,q} \dots R_{2,c,q} R_{1,c,q}$ is a proper prefix of $V_{w,q,q}$.*

Proof. This can easily be shown with an induction over w. ◀

▶ **Corollary 10.** *For $c \leq q-1$ the string $C_{v,c,q}$ is a proper substring of $V_{v,q,q}$*

This corollary leads to an upper bound for the highest power as well as for the necessary number of LZ77-factors of $S_{v,d,q}$.

▶ **Lemma 11.** *The string $S_{v,d,q}$ does not contain a $(2q+1)$-th power.*

Proof. by contradiction:

Assume there is a $q+1$-th power P in $S_{v,d,q}$.

The power P can not contain a \$ because the character \$ occurs only $d \leq q$ times in $S_{v,d,q}$. Therefore, using the previous lemma, the power P has to be a substring of $V_{v,q,q}$.

The power P can not contain a σ_v because the character σ_v occurs only once in $V_{v,q,q}$. Therefore the power P has to be a substring of $(V_{v-1,q,q})^q$.

The power P can not contain a σ_{v-1} because the character σ_{v-1} occurs only q times in $(V_{v-1,q,q})^q = ((V_{v-2,q,q})^q v_{v-1} (V_{v-2,q,q})^q)^q$. Therefore the power P has to be a substring of $(V_{v-2,q,q})^{2q}$.

It can be inductively shown that P can not contain σ_j for $j \in \{v-2, v-3, \dots, 1\}$ because the character σ_j occurs only $2q$ times in $(V_{j,q,q})^{2q} = ((V_{j-1,q,q})^q v_j (V_{j-1,q,q})^q)^{2q}$. Therefore the power P has to be a substring of $(V_{j-1,q,q})^{2q}$.

Since there are no characters left, this is a contradiction.

Therefore the string $S_{v,d,q}$ does not contain a $(2q+1)$-th power. ◀

▶ **Lemma 12.** *The string $S_{v,d,q}$ can be written with at most $1+3v+2d$ LZ77-factors.*

Proof. Since the string $V_{0,*,*}$ consist of a single letter, it can be written with a single LZ77-factor. By induction, the string $V_{v,c,q} = V_{v-1,q,q} \cdot (V_{v-1,q,q})^{c-1} \cdot \sigma_v \cdot (V_{v-1,q,q})^c$ can be written with at most $1+3v$ LZ77-factors. Using Corollary 10 yields that the string $S_{v,d,q} := V_{v,q,q} \cdot \left(\prod_{i=1}^{d} \$ \cdot C_{v,q-i,q} \right)$ can be written with at most $1+3v+2d$ LZ77-factors. ◀

In order to give a lower bound of the maximal repeats of $S_{v,d,q}$, we show that for natural numbers w, l, m and r with $1 \leq w \leq v-1$ and $1 \leq m+1 \leq q-d \leq l, r \leq q-1$

$$M_{w,l,m,r,q} := L_{1,l,q} L_{2,l,q} \dots L_{w,l,q} (V_{w,q,q})^m R_{w,r,q} \dots R_{2,r,q} R_{1,r,q}.$$

are maximal repeats of $S_{v,d,q}$.

▶ **Lemma 13.** *For*

$$\begin{aligned} w &\leq v-1 \\ m+1 \leq l, r &\leq q-1 \end{aligned}$$

the string $M_{w,l,m,r,q}$ is a proper prefix of $C_{v,l,q}$ and a proper suffix of $C_{v,r,q}$.

Proof. Using Lemma 9 the string $(V_{w,q,q})^m R_{w,r,q} \dots R_{2,r,q} R_{1,r,q}$ is a prefix of $(V_{w,q,q})^{m+1}$ which is a proper prefix of $L_{w+1,l,q}$. Therefore $M_{w,l,m,r,q}$ is a proper prefix of $C_{v,l,q}$. Similarly $M_{w,l,m,r,q}$ is a proper suffix of $C_{v,r,q}$. ◀

▶ **Corollary 14.** *If $1 \leq w \leq v-1$ and $1 \leq m+1 \leq q-d \leq l, r \leq q-1$ hold, the string $M_{w,l,m,r,q}$ is a maximal repeat of $S_{v,d,q}$.*

Proof. Since $M_{w,l,m,r,q}$ is a proper prefix of $C_{v,l,q}$, the string $\$M_{w,l,m,r,q}\sigma_*$ appears in $S_{v,d,q}$. Since $M_{w,l,m,r,q}$ is a proper prefix of $C_{v,r,q}$, the string $\sigma_* M_{w,l,m,r,q}\$$ appears in $S_{v,d,q}$. These two occurrences form a maximal pair. Therefore, the string $M_{w,l,m,r,q}$ is a maximal repeat of $S_{v,d,q}$. ◀

▶ **Corollary 15.** *The string $S_{v,d,q}$ has at least $(v-1)(q-d)d^2$ maximal repeats*

Combining Lemma 11, Lemma 12 and Corollary 15 yields Theorem 2 as given in the introduction:

▶ **Theorem 2.** *For $2000 \leq z \leq q$ there is a string S without q-th powers which can be expressed by z LZ77-factors and which has at least $\frac{1}{500} q z^3$ maximal repeats.*

Proof. Define $S = S_{\lfloor \frac{z}{9} \rfloor, \lfloor \frac{z}{3} \rfloor - 1, \lfloor \frac{q-1}{2} \rfloor}$. Using Lemma 11 the string S has no q-th power. Using Lemma 12 the string S can be described with $1 + 3\lfloor \frac{z}{9} \rfloor + 2\left(\lfloor \frac{z}{3} \rfloor - 1\right) \leq z$ LZ77-factors. Using Corollary 15 the string S has at least

$$\begin{aligned} &\left(\left\lfloor \frac{z}{9} \right\rfloor - 1\right)\left(\left\lfloor \frac{q-1}{2} \right\rfloor - \left(\left\lfloor \frac{z}{3} \right\rfloor - 1\right)\right)\left(\left\lfloor \frac{z}{3} \right\rfloor - 1\right)^2 \\ &\quad \geq \left(\frac{z}{9} - 2\right)\left(\frac{q}{2} - \frac{3}{2} - \frac{z}{3}\right)\left(\frac{z}{3} - 2\right)^2 \\ &\quad \geq \left(\frac{z}{9} - 2\frac{z}{2000}\right)\left(\frac{q}{2} - \frac{3}{2}\frac{q}{2000} - \frac{z}{3}\cdot\frac{q}{z}\right)\left(\frac{z}{3} - 2\frac{z}{2000}\right)^2 \\ &\quad = \left(\frac{1}{9} - 2\frac{1}{2000}\right)\left(\frac{1}{2} - \frac{3}{2}\frac{1}{2000} - \frac{1}{3}\right)\left(\frac{1}{3} - 2\frac{1}{2000}\right)^2 q z^3 \\ &\quad \geq \frac{1}{500} q z^3 \end{aligned}$$

maximal repeats. ◀

5 Conclusion

Since Theorem 1 suggests that well-compressed strings with many maximal repeats also have high powers and Theorem 8 even suggests that these high powers are not hidden inside the maximal repeats but are either a prefix or a suffix of them, it seems promising to do some more research on the maximal repeats of strings with high powers.

It might be possible to derive a data structure from the CDAWG by merging nodes stemming from similar powers of the same base. This data structure and its size as well as its usability will be determined in future work.

There are three more problems which should be researched:

The upper bound for the number of maximal repeats and the maximal repeats of the string given in section 4 differ by a factor of almost 1500. Even for strings with very high powers the factor is almost 500. This huge gap leaves room for further investigation.

The string in section 4 uses that the highest power is bigger than the parameter d. If the highest power is smaller than the number of LZ77-factors, the number of maximal repeats is only cp^3z for some constant c. It is an open question, whether the upper bound given by Theorem 1 is still tight up to constant for strings without high powers.

While the upper bound for the number of maximal repeats $3q(z+1)^3$ presented in this paper is tight up to a constant factor, the string $\sigma_1\sigma_2\ldots\sigma_{z-2}(\sigma_{z-1})^{q-1}$ has z LZ77-factors, no q-th power but a $(q-1)$-power and has only the $q-2$ maximal repeats $(\sigma_{z-1})^i$ with $1 \leq i \leq q-2$. Therefore, some additional structures should be taken into account in order to get a good estimate for the number of maximal repeats in a string.

References

1 Djamal Belazzougui, Fabio Cunial, Travis Gagie, Nicola Prezza, and Mathieu Raffinot. Composite Repetition-Aware Data Structures. In Ferdinando Cicalese, Ely Porat, and Ugo Vaccaro, editors, *Combinatorial Pattern Matching - 26th Annual Symposium, CPM 2015, Ischia Island, Italy, June 29 - July 1, 2015, Proceedings*, volume 9133 of *Lecture Notes in Computer Science*, pages 26–39. Springer, 2015. doi:10.1007/978-3-319-19929-0_3.

2 Anselm Blumer, J. Blumer, David Haussler, Ross M. McConnell, and Andrzej Ehrenfeucht. Complete inverted files for efficient text retrieval and analysis. *J. ACM*, 34(3):578–595, 1987. doi:10.1145/28869.28873.

3 Anselm Blumer, Andrzej Ehrenfeucht, and David Haussler. Average sizes of suffix trees and DAWGs. *Discrete Applied Mathematics*, 24(1-3):37–45, 1989. doi:10.1016/0166-218X(92)90270-K.

4 Moses Charikar, Eric Lehman, Ding Liu, Rina Panigrahy, Manoj Prabhakaran, Amit Sahai, and Abhi Shelat. The smallest grammar problem. *IEEE Trans. Information Theory*, 51(7):2554–2576, 2005. doi:10.1109/TIT.2005.850116.

5 Maxime Crochemore and Wojciech Rytter. *Text Algorithms*. Oxford University Press, 1994. URL: http://www-igm.univ-mlv.fr/%7Emac/REC/B1.html.

6 Chiara Epifanio, Filippo Mignosi, Jeffrey Shallit, and Ilaria Venturini. On Sturmian graphs. *Discrete Applied Mathematics*, 155(8):1014–1030, 2007. doi:10.1016/j.dam.2006.11.003.

7 N. J. Fine and H. S. Wilf. Uniqueness Theorems for Periodic Functions. *Proceedings of the American Mathematical Society*, 16(1):109–114, 1965. URL: http://www.jstor.org/stable/2034009.

8 Isamu Furuya, Takuya Takagi, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Takuya Kida. MR-RePair: Grammar Compression based on Maximal Repeats. *CoRR*, abs/1811.04596, 2018. arXiv:1811.04596.

9 Dan Gusfield. *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology*. Cambridge University Press, 1997.

10 Jacques Nicolas, Christine Rousseau, Anne Siegel, Pierre Peterlongo, François Coste, Patrick Durand, Sebastien Tempel, Anne-Sophie Valin, and Frédéric Mahé. Local and Maximal Repeats. URL: https://www.researchgate.net/publication/228940275_Local_and_Maximal_Repeats.

11 Jakub Radoszewski and Wojciech Rytter. On the structure of compacted subword graphs of Thue-Morse words and their applications. *J. Discrete Algorithms*, 11:15–24, 2012. doi:10.1016/j.jda.2011.01.001.

12 Mathieu Raffinot. On maximal repeats in strings. *Inf. Process. Lett.*, 80(3):165–169, 2001. doi:10.1016/S0020-0190(01)00152-1.

13 Yuka Tanimura, Takaaki Nishimoto, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda. Small-Space LCE Data Structure with Constant-Time Queries. In Kim G. Larsen, Hans L. Bodlaender, and Jean-François Raskin, editors, *42nd International Symposium on Mathematical Foundations of Computer Science, MFCS 2017, August 21-25, 2017 - Aalborg, Denmark*, volume 83 of *LIPIcs*, pages 10:1–10:15. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017. doi:10.4230/LIPIcs.MFCS.2017.10.