How Many Squares Can a String Contain?

Aviezri S. Fraenkel[1] and Jamie Simpson[2]

Abstract. All our words (strings) are over a *fixed* alphabet. A square is a subword of the form $uu = u^2$, where u is a nonempty word. Two squares are *distinct* if they are of different shape, not just translates of each other. A word u is *primitive* if u cannot be written in the form $u = v^j$ for some $j \geq 2$. A square u^2 with u primitive is *primitive rooted.* Let $M(n)$ denote the maximum number of distinct squares, $P(n)$ the number of distinct primitive rooted squares in a word of length n. We prove: no position in any word can be the beginning of the rightmost occurrence of more than two squares, from which we deduce $M(n) < 2n$ for all $n > 0$, and $P(n) = n - o(n)$ for infinitely many n.

1. Introduction

We consider words (strings, sequences) over a *fixed* alphabet throughout. A *square* is a pair of identical adjacent subwords, such as $1011010110 = 10110^2$ over the binary alphabet. Two squares are *distinct* if they are of different shape, not just translates of each other. Denote by $M(n)$ the maximum number of distinct squares occurring in a word of length n, the maximum being taken over all words of length n. A word u is *primitive* if u cannot be written in the form $u = v^j$ for some integer $j \geq 2$. Otherwise u is *nonprimitive*. The squares u^2 in the sequel do not necessarily have the property that u is primitive, unless explicitly stated so. For the case of squares u^2 such that u is not necessarily primitive, we consider distinct squares, as the problem of computing the number of repeated squares becomes trivial if we count repetitions of a square: the word 0^n has $n^2/4$ or $(n^2-1)/4$ repeated squares as n is, respectively, even or odd. This is clearly the maximum number of squares, since *every* even length block is a square.

It is clear that $M(n) \geq \lfloor n/2 \rfloor$ for all n, since this bound is achieved by the word 0^n. Values of $M(n)$ for binary squares and low n have been calculated and are shown in Table 1 below. The rightmost column gives the number of distinct words each of which has the corresponding maximum number of squares.

Our first main result states that no position in a string can be the beginning position of the rightmost occurrence of more than two squares. To state this precisely, we introduce a little notation. Consider the word $a_1 a_2 \cdots a_n$. For $i = 1, \ldots, n$ we let s_i be the number of squares beginning at

[1]Department of Applied Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 76100, Israel; `http://www.wisdom.weizmann.ac.il/~fraenkel` `fraenkel@wisdom.weizmann.ac.il`

[2]School of Mathematics, Curtin University, Perth WA 6001, Australia; `http://www.cs.curtin.edu.au/~simpson` `simpson@cs.curtin.edu.au`

Table 1.

n	Max number $M(n)$ of squares	Example	Number of maximal words
1	0	0	2
2	1	00	2
3	1	000	4
4	2	0000	4
5	2	00000	20
6	3	000000	12
7	3	0000000	66
8	4	00000000	46
9	5	001001010	20
10	6	0010110110	12
11	7	00110110101	4
12	7	000010110110	36
13	8	0000110110101	48
14	9		
16	11		
18	12		
20	13		
22	15		

position i which do not appear to the right of i. For example, in the word 001001010 we have $s_1 = 1$ since 001001 is a square beginning with a_1 which doesn't appear later on. The square 00 also begins with a_1 but this does not affect the value of s_1 since 00 appears later in the word. The other values are $s_2 = 1$, $s_3 = 0$, $s_4 = 1$, $s_5 = 1$, $s_6 = 1$, $s_7 = 0$, $s_8 = 0$, $s_9 = 0$.

From this example one might suspect that s_i can only take the values 0 and 1 but this is not the case. In the word 0100101001 we have $s_1 = 2$. If one might suspect that $s_1 = 2$ implies $s_2 \leq 1$, then the word 00100001001000010000100100 shows that we can have $s_1 = 2$ (with $(00100001)^2, (0010000100100)^2$) and $s_2 = 2$ (with $(01000010)^2, (0100001001000)^2$) and $s_3 = 1$ (with $(10000100)^2$).

In §2 we prove,

Theorem 1. *For every nonempty word of length n, $s_i \leq 2$ for all $i \in \{1, \ldots, n\}$.*

Theorem 1 implies,

Corollary 1. *For an arbitrary alphabet, $M(n) \leq 2n-8$ for $n \geq 5$; and $M(n)$ is given by Table 1 for $n \in \{1, \ldots, 5\}$.*

Corollary 2. *For the binary alphabet, $M(n) \leq 2n-29$ for $n \geq 22$; and $M(n)$ is given by Table 1 for $n \in \{1, \ldots, 22\}$.*

In §3 we prove a lower bound result, namely,

Theorem 2. *For infinitely many positive integers n we have $M(n) = n - o(n)$.*

In fact, we show that Theorem 2 holds even for a binary alphabet. For other alphabets, larger upper bounds may apply.

In a complementary paper [**FrS1995**], we established the *minimum* number of distinct squares that can exist in a *binary* word. A binary word containing no square has maximum length 3; in fact, the only binary squarefree words of length 3 are 010 and its 1-complement 101. Let $g(k)$ denote the length of a longest binary word containing at most k distinct squares. Thus $g(0) = 3$. A computer disclosed that $g(1) = 7$, $g(2) = 18$. The main result of the complementary paper was to show that $g(3) = \infty$, i.e., there is an arbitrarily long binary word containing only 3 squares; such a word, containing only the squares $u^2 = 0^2$, 1^2 and $(01)^2$, was constructed there. Note that each of the three u is primitive. Previously it had been shown that $g(5) = \infty$; see [**EJS1974**].

A square u^2 with u primitive will be called a *primitive rooted* square. Denote by $P(n)$ the maximum number of distinct primitive rooted squares in a word of length n. The method of proof of Theorem 2 permits us to show the following result:

Corollary 3. *Similarly to Corollaries* 1 *and* 2*, for an arbitrary fixed alphabet, $P(n) \le 2n-8$ for $n \ge 5$; for the binary alphabet, $P(n) \le 2n - 29$ for $n \ge 22$. Further, $P(n) = n - o(n)$ for infinitely many n.*

Fibonacci words are defined by $f_0 = 0$, $f_1 = 1$, $f_m = f_{m-1}f_{m-2}$ (concatenation) for all $m \ge 2$. Let $F_m = |f_m|$. Then $F_0 = 1$, $F_1 = 1$, $F_m = F_{m-1} + F_{m-2}$ $(m \ge 2)$ are the Fibonacci numbers, since $|f_m| = |f_{m-1}f_{m-2}| = |f_{m-1}| + |f_{m-2}|$. Note that all squares in f_m $(m \ge 0)$ are primitive rooted (see e.g., [**CrR1994**], §2). Denote by $D(n)$ and by $R(n)$ the exact number of *D*istinct and *R*epeated squares respectively in the Fibonacci word f_n. As usual, let $\varphi = (1+\sqrt{5})/2$ denote the golden section (so $\varphi^2 = \varphi + 1$), and denote by $N(x)$ the nearest integer to the real number x. In [**FrS≥1997**] it was shown (Theorem 1) that for $n \ge 5$, $D(n) = 2(F_{n-2} - 1)$, leading to (Corollary 1 there): $D(n) = 2(2-\varphi)F_n + o(1)$ $\quad (2(2-\varphi) \approx 0.7639)$, and also $D(n) = N\left(\frac{2\varphi^{n-1}}{\sqrt{5}} - 2\right)$. From the asymptotic formula for $D(n)$ we get, $M(n) \ge 2(2-\varphi)n$ for infinitely many n. Thus we see that Fibonacci words do not contain as many distinct squares as obtained with the special set of words used in §3 below.

Fibonacci words have been used for giving lower bounds on the number of repeated primitive squares. In [**Cro1981**], Lemma 10, it was shown that $R(n) \ge \frac{1}{6}F_n \log_2 F_n$ for $n \ge 5$. Theorem 11 in [**CrR1995**] asserts that the number of square prefixes of a word w is $< \log_\phi |w|$. For $w = f_n$ this implies $R(n) < F_n \log_\phi F_n = F_n \log_2 F_n / \log_2 \phi \le 1.441 F_n \log_2 F_n$. In [**FrS≥1997**] an exact formula for $R(n)$ is derived, which implies: $R(n) = \frac{2}{5}(3-\varphi)nF_n + O(F_n)$ $\quad (\frac{2}{5}(3-\varphi) \approx 0.7962)$. We don't know whether there is a class of words which has a number of repeated primitive squares larger than Fibonacci words.

2. The Upper Bound

For establishing Theorem 1 we prove the following lemma.

Lemma 1. *Let u^2, v^2, w^2 be three prefixes of a word x such that w is primitive and $p > q > r$, where $p = |u|$, $q = |v|$, $r = |w|$. Then $p \ge q + r$.*

Proof. If u, v, w are all primitive, the result is Lemma 10 of [**CrR1995**]. Examining their proof we see that actually only the primitivity of w is used. ∎

Remarks.

(a) If w is nonprimitive, the result doesn't hold in general. Thus $u = 0^4$, $v = 0^3$, $w = 0^2$ is a counterexample for every word with prefix 0^8.

(b) In [**CrR1995**] the result is stated in the form $p > q + r$, but only $p \geq q + r$ is true and proved. In fact, for 0100101001001010 we can take $u = 01001010$, $v = 01001$ and $w = 010$, all of which are primitive, yet $|u| = |v| + |w|$.

The following property of the function s_i, introduced in §1, is immediate.

For each square, some occurrence of it (the rightmost) adds one to the value of s_i for some i. Hence we have $\sum_{i=1}^{n} s_i \leq M(n)$; and if x is a word of length n which achieves the maximum of $M(n)$ squares, then

$$\sum_{i=1}^{n} s_i = M(n) \tag{1}$$

for x. Incidentally, since we cannot have a square beginning with the last letter of a word, we have $s_n = 0$ for every word of length n.

A *period* of a word $x = a_1 a_2 \ldots a_n$ is any integer p in $\{1, \ldots, n\}$ for which $a_i = a_{i+p}$ for all $i \in \{1, \ldots, n - p\}$. The following is the weak form of the Periodicity Lemma of Fine and Wilf [**FiW1965**]. See also [**Lot1983**], [**CrR1994**].

Periodicity Lemma. *Let p and q be two periods of a word x. If $p + q \leq |x|$, then* $\gcd(p, q)$ *is also a period of x.*

Proof of Theorem 1. We prove this by contradiction. Suppose that there exists a word x of length n and some $i \in \{1, \ldots, n - 1\}$ for which $s_i \geq 3$. Then $x = a_i a_{i+1} \ldots a_n$ has three prefixes u^2, v^2, w^2 which do not occur elsewhere in the word, and for which, without loss of generality, $p > q > r$, where $p = |u|$, $q = |v|$ and $r = |w|$. If $p \geq 2r$, then w^2 occurs again at position $p + 1$, because u occurs there — a contradiction. Thus $p < 2r$, so

$$p > q > r > \frac{p}{2}. \tag{4}$$

By (4), $p < 2r < q + r$. It follows from Lemma 1 that w is nonprimitive, so there exists a primitive y such that $w = y^t$ for some integer $t \geq 2$. Let $r_1 = |y|$, so $r = tr_1$. Then u^2, v^2 and y^2 are prefixes of x with y primitive, so by Lemma 1,

$$r_1 + q \leq p\,. \tag{5}$$

Note that r_1 is a period of w^2; it is also a period of u, since $p < 2r$ (by (4)). Also q is a period of u since $q < p < 2q$ by (4), so u is a prefix of v^2. Using (5) we can therefore apply the Periodicity Lemma to conclude that $\gcd(r_1, q) = d$ is also a period of u. Then $d | r_1$ and $d | q$. Since y is primitive, we have $d = r_1$. Thus $r_1 | q$. Now $r = tr_1$, so $q = (t + s) r_1$ for some integer $s \geq 1$. Therefore v^2 has length $2(t + s) r_1$ and w^2 has length $2tr_1$. It follows that w^2 appears also at $r_1 + 1$, which is a contradiction. ■

Proof of Corollary 1. Since $s_i \leq 2$ for all i (Theorem 1) and $s_n = 0$, we conclude that $M(n) < 2n$. It is easy to verify that the three left columns of Table 1 are valid for any fixed alphabet for $n \leq 5$. Thus the tail end of length 5 of any word of length $ge5$ contains at most 2 squares. Such a word has thus at most $2(n - 5) + 2 = 2n - 8$ squares. ■

Proof of Corollary 2. The same argument as in the preceding proof applies, but Table 1 now shows that the tail end of length 22 of any binary word of length $ge22$ contains at most 15 squares. Such a word has thus at most $2(n-22)+15=2n-29$ squares. ■

3. The Lower Bound

We construct a class of words realizing the lower bound enunciated in Theorem 2. For examining the properties of these words, we prove two lemmas.

For $i \geq 1$, let $q(i)$ be the word $0^{i+1}10^i10^{i+1}1$. Thus $q(1)=00101001$, $q(2)=00010010001$, etc. For $m \geq 1$, let $Q=Q(1,m)$ be the concatenation $q(1)q(2)\cdots q(m)$.

Lemma 2. *Q contains $(3m^2+13m)/2$ bits.*

Proof. Clearly $q(i)$ contains $3i+5$ bits, so $Q=\sum_{i=1}^{m}(3i+5)=(3m^2+13m)/2$. ■

Lemma 3. *Q contains*

$$n = 3m^2/2 + \lfloor (m+1)/2 \rfloor + 7m/2 - 3 \tag{6}$$

squares.

Proof. Denote the number of squares in Q which contain 0, 2 and 4 1s by n_0, n_2 and n_4 respectively. We calculate each of these and then show that the word contains no other squares. The total number of squares is then the sum of these three amounts.
We have

$$n_0 = \lfloor (m+1)/2 \rfloor,$$

since there are $m+1$ consecutive 0s in $q(m)$ and there is no longer block of consecutive 0s in Q.

Now $q(i)$ contains the $i+1$ squares $0^i1\,0^i1, 0^{i-1}10\,0^{i-1}10, \ldots, 10^i\,10^i$. That is, we get all possible squares containing 2 1s and $2i+2$ bits. Note that the same squares, except the first and last, are also obtained in the concatenation $q(i-2)q(i-1)$. Therefore there are m additional squares containing 2 1s of length $2m+4$ in $q(m-1)q(m)$. These are

$$0^m10\,0^m10, 0^{m-1}100\,0^{m-1}100, \ldots, 010^m\,010^m.$$

Thus Q contains precisely $2+3+\cdots+(m+1)+m$ such squares, which implies

$$n_2 = m^2/2 + 5m/2.$$

We now count the squares containing 4 1s. In each triple

$$q(i)q(i+1)q(i+2) = 0^{i+1}10^i10^{i+1}1\,0^{i+2}10^{i+1}10^{i+2}1\,0^{i+3}10^{i+2}10^{i+3}1\,,$$

there are $2i+3$ such squares containing $4i+10$ bits. The first of these begins at the bit in position $i+3$ and the last at position $3i+5$. They have the form

$$0^i10^{i+1}10^2\,0^i10^{i+1}10^2, 0^{i-1}10^{i+1}10^3\,0^{i-1}10^{i+1}10^3, \ldots, 10^{i+2}10^{i+1}\,10^{i+2}10^{i+1}\,.$$

Note that each begins in the subword $q(i)$ and that these are the only squares containing 4 1s which begin in $q(i)$. We have $m-2$ such triples, with i ranging from 1 to $m-2$ so we get $\sum_{i=1}^{m-2}(2i+3)=m^2-4$ such squares. We also get $m+1$ squares containing 4 1s which begin in $q(m-1)$ and are contained in

$$q(m-1)q(m) = 0^m10^{m-1}10^m1\,0^{m+1}10^m10^{m+1}1\,.$$

These contain $4m+6$ bits. The first begins at the bit at position $m+2$ and the last at the bit at position $2m+2$. The first is $0^{m-1}10^m10^2\,0^{m-1}10^m10^2$ and the last is $0^m10^{m+1}1\,0^m10^{m+1}1$. It is clear that these are the only squares containing 4 1s which begin in $q(m-1)$ and that no square containing 4 1s can begin in $q(m)$; so we have found all squares in Q which contain 4 1s. Thus

$$n_4 = m^2 - 4 + m + 1 = m^2 + m - 3\,.$$

Finally we need to show there are no squares in the word containing more than 4 1s. Each half of such a square would contain a subword 10^x10^y1, but it is not hard to see that such words are unique in Q except that $q(i)q(i+1) = 0^{i+1}10^i10^{i+1}1\,0^{i+2}10^{i+1}10^{i+2}1$ contains twin appearances of the word $10^{i+1}10^{i+2}1$, but they share a common 1-bit (the first 1-bit in $q(i+1)$). Therefore they cannot be in adjacent halves of a square. Thus Q contains only squares with 4 or fewer 1s and the number of squares is

$$n = n_0 + n_2 + n_4 = 3m^2/2 + \lfloor (m+1)/2 \rfloor + 7m/2 - 3\,,$$

as required. ■

Proof of Theorem 2. The difference between the length $(3m^2+13m)/2$ of Q and the number n of the form (6) of squares it contains, is $3m - \lfloor (m+1)/2 \rfloor + 3$. For every $\varepsilon > 0$, this difference is $< \varepsilon n$ for all sufficiently large n of the form (6). ■

Proof of Corollary 3. The upper bound carries over from Corollaries 1 and 2 for $n \geq 5$ and $n \geq 22$ respectively, since $P(n) \leq M(n)$, which follows directly from the definition of $P(n)$ and $M(n)$. As to the lower bound, note that the squares in the word Q defined at the beginning of this section are all primitive rooted, except for the blocks of length > 2 of consecutive 0s. But there is only a linear number of those, namely, $\lfloor (m+1)/2 \rfloor - 1$. Their deletion leaves the lower bound of Theorem 2 invariant. ■

Remark. It is actually clear that $P(n) = M(n)$ for $n \in \{1, \ldots, 5\}$ for any fixed alphabet; and Table 2 shows that $P(n) = M(n)$ for $n \in \{1, \ldots, 13\}$ for the binary case.

We end with some natural questions.

(a) Close the gap between Corollaries 1, 2 and Theorem 2. In particular, is it true that for every fixed alphabet, $M(n) < n$ for all $n > 0$?
(b) Is there an upper bound for $P(n)$, better than that implied by Corollary 1 or Corollary 2?
(c) Is there a lower bound for either $M(n)$ or $P(n)$, better than that of Theorem 2 for a fixed *nonbinary* alphabet?
(d) Find upper and lower bounds for the *smallest* number of repeated nonprimitive/primitive rooted squares in a word. (The case of distinct nonprimitive/primitive rooted squares was solved in [**FrS1995**].)

Acknowledgments. During the International Workshop on Pattern Matching, Text Processing and Compression, University of Warwick, Coventry, UK, December 1996, we had interesting conversations about the subject matter with Alberto Apostolico, Maxime Crochemore and Filippo Mignosi. In particular, we learned about reference [CrR1995], Lemma 10 of which enabled us to shorten our original proof of Theorem 1. Dan Gusfield improved on what we believed to be our

TABLE 2

n	$P(n)$	Example
1	0	0
2	1	00
3	1	000
4	2	0011
5	2	00011
6	3	001011
7	3	0001011
8	4	00101011
9	5	001001010
10	6	0010110110
11	7	00110110101
12	7	001101101010
13	8	0011011010110

final version. Thanks to programmers Kim Shearer who calculated Table 1, and Michael Howard who used a genetic algorithm to find long words containing many squares.

References

1. [Cro1981] M. Crochemore, An optimal algorithm for computing the repetitions in a word, *Inform. Process. Letters* **12** (1981) 244–250.

2. [CrR1994] M. Crochemore and W. Rytter, *Text Algorithms*, Oxford University Press, New York, 1994.

3. [CrR1995] M. Crochemore and W. Rytter, Squares, cubes, and time-space efficient string searching, *Algorithmica* **13** (1995) 405–425.

4. [EJS1974] R. Entringer, D. Jackson and J. Schatz, On nonrepetitive sequences, *J. Combin. Theory* (Ser. A) **16** (1974) 159–164.

5. [FiW1965] N.J. Fine and H.S. Wilf, Uniqueness theorems for periodic functions, *Proc. AMS* **16** (1965) 109–114.

6. [FrS1995] A.S. Fraenkel and R.J. Simpson, How many squares must a binary sequence contain? *Electronic J. Combinatorics* **2** (1995) R2, 9pp.
http://www.combinatorics.org/Journal/journalhome.html

7. [FrS≥1997] A.S. Fraenkel and R.J. Simpson, The Exact Number of Squares in Fibonacci Words, preprint.

8. [Lot1983] M. Lothaire, *Combinatorics on Words*, Addison-Wesley, Reading, MA, 1983.