Three Overlapping Squares: The General Case Characterized*

W. F. Smyth[1,2]

[1] Algorithms Research Group, Department of Computing & Software
McMaster University, Hamilton, Ontario, Canada L8S 4K1
smyth@mcmaster.ca

[2] Department of Mathematics & Statistics
University of Western Australia, Crawley WA 6009, Australia

Abstract. The "Three Squares Lemma" [7] famously explored the consequences of supposing that three squares occur at the same position in a string; essentially it showed that this phenomenon could not occur unless the longest of the three squares was at least the sum of the lengths of the other two. More recently, several papers [8, 25, 17, 11] have greatly extended this result by supposing that only two of the squares occur at the same position, with a third occurring in a neighbourhood to the right — in these cases also, similar restrictions apply. In this paper an alternative strategy is proposed: the consequences of having only *two* squares at neighbouring positions are carefully analyzed, and then the observation is made that the analysis applies in a straightforward way (though perhaps with complicated details) to the three neighbouring squares problem in its full generality. This problem and the methodology required to solve it are outlined. I conclude with a brief discussion of the potential application of this research to an entirely new approach to the computation of maximal periodicities in a string.

Keywords: string, word, overlapping squares, repetition, run, maximal periodicity.

1 Introduction

Beginning with the "Three Squares Lemma" of Crochemore & Rytter [7], there has for several years been considerable interest in the limitations that may exist on periodicity in strings. [7] showed that three squares could exist at the same position in a string only if the longest of the three was at least the sum of the lengths of the other two. A sequence of papers [8, 25, 17, 11] greatly generalized this result and also made it more precise by considering two squares $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$ at the same position, with however the third square $\boldsymbol{w}^2$ offset a distance $k \geq 0$

* This work was supported in part by the Natural Sciences & Engineering Research Council of Canada.

to the right. The analysis given in the cited papers deals with 12 of 14 subcases that arise: two remain to be considered, but it seems clear that such behaviour is impossible — that is, the assumption that three neighbouring squares of well-defined size exist within these well-defined bounds leads to the conclusion that locally the string breaks down into repetitions of small period. In this paper I show how to characterize the general case of overlapping squares — no two constrained to begin at the same position — and make a start on considering the combinatorial consequences.

Interest has been added to this research by a parallel development over the last dozen years or so: the attempt to specify sharp bounds on the number of maximal periodicities ("runs") that can occur in any string of given length n. Kolpakov & Kucherov [16] showed that the maximum number of runs (usually denoted $\rho(n)$) was linear in n, and moreover they described a linear-time algorithm to compute all the runs in any given string; but their proof was nonconstructive — the maximum number of runs was shown to be $\Theta(n)$ but no constant of proportionality was specified. As briefly described in Section 2, the resulting research, some combinatorics and much computing, has led to the conclusion that $\rho(n)$ is at least $0.944575n$ [26] and at most $1.029n$ [6] — in other words, more or less the string length n. What links these two streams of research is a simple observation:

> If the maximum number of runs over all strings of length n is itself approximately n, then on average there will be about one run starting at each position. Thus, if two runs start at some position, there must be some other position, probably nearby, at which no run can start. More generally, determining combinatorial constraints on the occurrence of overlapping squares (runs) may lead to a better characterization of $\rho(n)$.

There is a third avenue of research that relates closely to overlapping squares: the computation of all the runs/repetitions in a given string. At present the only way that this can be done is essentially by brute force: global data structures (suffix array, longest common prefix array, Lempel-Ziv decomposition) need to be computed in an extended preprocessing phase, when in fact it has been shown [21] that the *expected* number of runs in a string is generally much less than string length. The preprocessing is necessitated by the absence of a detailed understanding of the combinatorics of overlapping occurrences of runs in strings.

In Section 2 terminology and notation are introduced; Section 3 shows how to express the general case of three overlapping squares, together with a sample lemma that applies the insights gained; finally, in Section 4 I suggest possible future research directions.

2 Preliminaries

(Usage generally follows [27].) A ***string*** is a finite sequence of symbols (***letters***) drawn from some finite or infinite set Σ called the ***alphabet***. The alphabet ***size*** is $\sigma = |\Sigma|$. We write a string $\boldsymbol{x}$ in mathbold, and we represent it as an array

$\boldsymbol{x}[1..n]$ for some $n \geq 0$. We call $n = x$ the ***length*** of $\boldsymbol{x}$. For $x = 0$, $\boldsymbol{x} = \boldsymbol{\varepsilon}$, the ***empty string***.

If $\boldsymbol{x} = \boldsymbol{uvw}$, then $\boldsymbol{u}$ is said to be a ***prefix***, $\boldsymbol{v}$ a ***substring*** and $\boldsymbol{w}$ a ***suffix*** of $\boldsymbol{x}$. If $\boldsymbol{x} = \boldsymbol{uv}$, $0 \leq u < x$, then $\boldsymbol{vu}$ is said to be the $u^{\boldsymbol{th}}$ ***rotation*** of $\boldsymbol{x}$, written $R_u(\boldsymbol{x})$. If $\boldsymbol{x} = \boldsymbol{uv} = \boldsymbol{wu}$ for $u < x$, then $\boldsymbol{u}$ is a ***border*** of $\boldsymbol{x}$, and $\boldsymbol{x}$ has ***period*** $p = x - u$; that is, for every $i \in 1..u$, $\boldsymbol{x}[i] = \boldsymbol{x}[i+p]$. The string

$$\begin{array}{cccccccccccc} & & {\scriptstyle 1} & {\scriptstyle 2} & {\scriptstyle 3} & {\scriptstyle 4} & {\scriptstyle 5} & {\scriptstyle 6} & {\scriptstyle 7} & {\scriptstyle 8} & {\scriptstyle 9} & {\scriptstyle 10} \\ \boldsymbol{x} & = & a & b & a & a & b & a & b & a & a & b \end{array} \tag{1}$$

has borders $abaab$ and ab, hence corresponding periods 5 and 8, respectively.

If $\boldsymbol{x} = \boldsymbol{vu}^e\boldsymbol{w}$, where $e \geq 2$ and $\boldsymbol{u}$ is neither a suffix of $\boldsymbol{v}$ nor a prefix of $\boldsymbol{w}$ (the integer e is maximum), then $\boldsymbol{u}^e$ is said to be a ***repetition*** in $\boldsymbol{x}$. The integers u and e are the ***period*** and ***exponent***, respectively, of the repetition. The string (1) has repetitions $(aba)^2, (abaab)^2, a^2, (ab)^2, (ba)^2$, each of which is a ***square***. In general, every repetition has a square prefix.

If $\boldsymbol{v} = \boldsymbol{x}[i..j]$ has period u, where $v/u \geq 2$, and if neither $\boldsymbol{x}[i{-}1..j]$ nor $\boldsymbol{x}[i..j{+}1]$ (whenever these are defined) has period u, then $\boldsymbol{v}$ is said to be a ***maximal periodicity*** or ***run*** in $\boldsymbol{x}$ [18] with a (now fractional) ***exponent*** $e = v/u$. All of the repetitions in (1) are runs except for $(ab)^2$ and $(ba)^2$: these are substrings of the run $\boldsymbol{v} = ababa$. In general, every repetition is a substring of some run; thus computing all the runs implicitly computes all the repetitions.

There were three classical algorithms proposed [3, 1, 19] for computing all the repetitions in a string of length n, each executing in $O(n \log n)$ time, asymptotically optimal since the ***Fibonacci string $\boldsymbol{f_k}$***, defined by

$$\boldsymbol{f_0} = b, \boldsymbol{f_1} = a;\ k \geq 2 \Longrightarrow \boldsymbol{f_k} = \boldsymbol{f_{k-1}f_{k-2}},$$

contains $O(f_k \log f_k)$ repetitions [3, 15, 10]. In [18] Main proposed an algorithm to compute all the "leftmost" runs, extended by Kolpakov & Kucherov in [16] to compute all runs. As mentioned in Section 1, this approach makes extensive use of preprocessing, but still executes in linear time, based on a complex proof that the maximum number $\rho(n)$ of runs in any string of length n satisfies

$$\rho(n) \leq K_1 n - K_2 \sqrt{n} \log_2 n \tag{2}$$

for some universal positive constants K_1 and K_2. Even though [16] provided computational evidence (up to $n = 60$) that $\rho(n) \leq n$, the method of proof allowed no bounds to be placed on K_1 and K_2. Over the last decade, the bounding of $\frac{\rho(n)}{n}$ has become a growth industry, leading to a lower bound 0.944575 [12, 20, 26] and an upper bound 1.029 [23, 22, 4, 13, 14, 5, 6], the latter result achieved using three years of CPU time on a supercomputer [24]. Meanwhile, more efficient algorithms for computing runs have been proposed — for example, [2] — but still with heavy preprocessing and the same general approach. Since, as noted in Section 1, runs are expected to be sparse in strings, even for small sigma [21], a heavy-handed global approach seems inappropriate.

A parallel approach has sought to find a combinatorial basis for estimating the maximum number of runs in a string, specifically by considering the consequences of assuming that two squares occur at the same position in a string, with a third nearby, somewhat to the right. This generalizes the "Three Squares Lemma" [7] that considered three squares at the same position in the string:

Lemma 1 *Suppose* $\boldsymbol{u}$ *is not a repetition, and suppose* $\boldsymbol{v} \neq \boldsymbol{u}^j$ *for any* $j \geq 1$*. If* $\boldsymbol{u}^2$ *is a prefix of* $\boldsymbol{v}^2$*, in turn a proper prefix of* $\boldsymbol{w}^2$*, then* $w \geq u{+}v$*.*

A series of papers, particularly [8, 25, 17, 11], has considered the following more general problem:

> **(P)** Suppose that a string $\boldsymbol{x}$ has prefixes $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$, $3u/2 < v < 2u$, and suppose further that a third square $\boldsymbol{w}^2$ occurs at position $k{+}1$ of $\boldsymbol{x}$, where $v{-}u < w < v$, $w \neq u$, and $0 \leq k < v{-}u$. What can be said about the periodicity of $\boldsymbol{x}$?

It turns out that the solution to Problem (P) breaks down into 14 subcases, depending on the relative sizes of the parameters k, u, v, w; of these 12 have been considered in detail and for each of them it turns out that $\boldsymbol{x}$ breaks down into repetitions of small period — essentially, the postulate of three such squares cannot be satisfied. There is good computational evidence that the remaining two subcases exhibit the same behaviour.

Moreover, it is clear that further generalization is of interest: what happens when the three squares $\boldsymbol{u}^2, \boldsymbol{v}^2, \boldsymbol{w}^2$ are merely constrained to be "neighbouring", without the requirement that $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$ occur at the same position? What is an appropriate formulation of such a problem? What relative values of k, u, v, w are of combinatorial interest?

In this paper I begin to answer these questions by first considering only two overlapping squares in some detail, then making the observation that three overlapping squares can always be thought of as two sets of two overlapping squares. In Section 3 a general lemma for two squares is stated and proved, and then a "sample" three squares lemma is proved, based on the characterization of the general case for two squares.

3 Characterizing the General Case

We are interested in the cases that arise when a square $\boldsymbol{u}^2$ beginning at some position i in a string overlaps with a second square $\boldsymbol{v}^2$ at position $i{+}k$, $k \geq 0$, to its right.

Lemma 2 *Suppose* $\boldsymbol{x}$ *has prefixes* $\boldsymbol{u}^2$ *and* $\boldsymbol{k}\boldsymbol{v}^2$*,* $k \geq 0$*, where* $x = \max(2u, k{+}2v)$*,* $k \leq u < 2v$*.*

(a) $k{+}v < u < 2v$ $(k < \min(v{-}1, u{-}v))$:

$$\boldsymbol{x} = (\boldsymbol{p}^e\boldsymbol{z})^2 = \boldsymbol{p}^e\boldsymbol{q}^f\boldsymbol{q}^{f-e} = \boldsymbol{p}^e\boldsymbol{q}^f\boldsymbol{p}[k{+}1..u{-}v],$$

where $\boldsymbol{p} = \boldsymbol{u}[1..u{-}v]$*,* $e = \frac{k+v}{u-v} > 1$*,* $\boldsymbol{z} = \boldsymbol{v}[1..u{-}(k{+}v)$*,* $\boldsymbol{q} = R_k(\boldsymbol{p})$*,* $f = \frac{u}{u-v} > 2$*,* $f{-}e \leq 1$*.*

(b) $\frac{k}{2}+v \le u \le k+v$ $(1 \le u-v \le k \le 2(u-v))$:

$$\boldsymbol{x} = (\boldsymbol{z}\boldsymbol{p}^e)^2 = (\boldsymbol{q}[1..k+v-u]\boldsymbol{p}^e)^2 = (\boldsymbol{k}\boldsymbol{p}^{e-1})^2,$$

where $\boldsymbol{z} = \boldsymbol{u}[1..k+v-u]$, $\boldsymbol{p} = \boldsymbol{u}[1..u-v]$, $e = 1+\frac{u-k}{u-v} \ge 1$, $\boldsymbol{q} = R_d(\boldsymbol{p})$, $d = (u-k) \bmod (u-v)$.

(c) $v < u < \frac{k}{2}+v$ $(k > 2(u-v))$:

$$\boldsymbol{x} = (\boldsymbol{q}\boldsymbol{y}\boldsymbol{p}^e)^2\boldsymbol{y},$$

where $\boldsymbol{p} = \boldsymbol{v}[1..u-v]$, $e = 1+\frac{u-k}{u-v} > 1$, $\boldsymbol{q} = R_d(\boldsymbol{p})$, $d = (u-k) \bmod (u-v)$, $\boldsymbol{y} = \boldsymbol{v}[2u-(k+v)+1..v]$. *Moreover, both* $\boldsymbol{x}$ *and* $\boldsymbol{k}\boldsymbol{v}$ *have border* $\boldsymbol{q}\boldsymbol{y}$.

(d) $\frac{2(k+v)}{3} \le u < v$ $(k \le \frac{3u}{2}-v < \frac{v}{2})$:

$$\boldsymbol{x} = (\boldsymbol{k}\boldsymbol{p}^e)^2\boldsymbol{q}\boldsymbol{k}\boldsymbol{p},$$

where $\boldsymbol{p} = \boldsymbol{v}[1..v-u]$, $e = \frac{u-k}{v-u} > 1$, $\boldsymbol{q} = R_d(\boldsymbol{p})$, $d = (u-k) \bmod (v-u)$. *Both* $\boldsymbol{x}$ *and* $\boldsymbol{k}\boldsymbol{v}$ *have border* $\boldsymbol{k}\boldsymbol{p}$.

(e) $\frac{k+v}{2} < u < \frac{2(k+v)}{3}$ $(\frac{3u-2v}{2} < k < 2u-v)$:

$$\boldsymbol{x} = \boldsymbol{k}(\boldsymbol{p}^e\boldsymbol{k}\boldsymbol{p})^2,$$

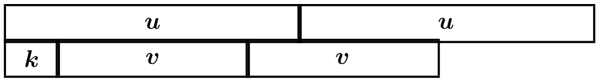
where $\boldsymbol{p} = \boldsymbol{v}[1..v-u]$, $e = \frac{u-k}{v-u} > 1$.

(f) $k \le u \le \frac{k+v}{2}$ $(\boldsymbol{u}^2$ *a prefix of* $\boldsymbol{k}\boldsymbol{v})$:

$$\boldsymbol{x} = \boldsymbol{k}(\boldsymbol{p}^e\boldsymbol{z})^2,$$

where $\boldsymbol{p} = \boldsymbol{u}[k+1..u]\boldsymbol{u}[1..k]$, $e = \frac{2u-k}{u} \ge 1$, $\boldsymbol{z} = \boldsymbol{v}[2u-k+1..v]$.

Proof. (a) Let $\boldsymbol{z} = \boldsymbol{u}[k+v+1..u] = \boldsymbol{v}[1..u-(k+v)]$, suffix of $\boldsymbol{u}$ and prefix of $\boldsymbol{v}$.



Observe that

$$\boldsymbol{u}[k+j] = \boldsymbol{v}[j] = \boldsymbol{u}[j-z], \ z+1 \le j \le v,$$

so that $\boldsymbol{u}[1..k+v] = \boldsymbol{k}\boldsymbol{v}$ has period $k+z = u-v = p$ (where $\boldsymbol{p} = \boldsymbol{u}[1..u-v]$). Consequently, we may write $\boldsymbol{x} = (\boldsymbol{p}^e\boldsymbol{z})^2$, where $e = \frac{k+v}{u-v} > 1$ (since $k+v < u$). Noting that $\boldsymbol{v} = \boldsymbol{u}[k+1..u-z]$, with $k < u-v$, we see also that $\boldsymbol{u} = \boldsymbol{k}\boldsymbol{q}^{f-1}\boldsymbol{z}$, where $\boldsymbol{q} = R_k(\boldsymbol{p})$,

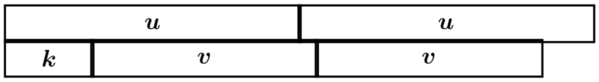
$$f = \frac{u}{u-v} = \frac{v}{u-v}+1 > 2.$$

Hence $\boldsymbol{x} = \boldsymbol{p}^e\boldsymbol{z}\boldsymbol{k}\boldsymbol{q}^{f-1}\boldsymbol{z}$. But $\boldsymbol{z}\boldsymbol{k}$ is a prefix of the second copy of $\boldsymbol{v}$ of length p, and comparing with the first copy of $\boldsymbol{v}$, we see that therefore $\boldsymbol{z}\boldsymbol{k} = R_k(\boldsymbol{p}) = \boldsymbol{q}$. Since moreover $\boldsymbol{z} = \boldsymbol{q}^g$, where

$$g = \frac{z}{u-v} = \frac{u}{u-v} - \frac{k+v}{u-v} = f-e \le 1,$$

we find $\boldsymbol{x} = \boldsymbol{p}^e\boldsymbol{q}^f\boldsymbol{q}^{f-e}$, as claimed. (Note that $g = 1$ iff $k = 0$.) Finally, writing $\boldsymbol{q} = \boldsymbol{p}[k+1..u-v]\boldsymbol{p}[1..k]$ and $f-e = 1-\frac{k}{u-v}$, we find that $\boldsymbol{q}^{f-e} = \boldsymbol{p}[k+1..u-v]$.

(b) Let $\boldsymbol{z} = \boldsymbol{u}[1..k+v-u]$, a possibly empty prefix of $\boldsymbol{u}$ and suffix of $\boldsymbol{v}$.



Observe that

$$\boldsymbol{u}[k+j] = \boldsymbol{v}[j] = \boldsymbol{u}[z+j],\ 1 \le j \le v-z = u-k,$$

where $z = k+v-u < k$. Thus $\boldsymbol{u}[z+1..u]$ and $\boldsymbol{v}$ have period $k-z = u-v = p$. Consequently, setting $\boldsymbol{p} = \boldsymbol{u}[z+1..k] = \boldsymbol{v}[1..k-z]$, we may write $\boldsymbol{x} = (\boldsymbol{z}\boldsymbol{p}^e)^2$, where $e = \frac{u-z}{u-v} \ge 1$, since $z \le v$. Noting that

$$u-z = u-k-v+u = (u-v)+(u-k), \tag{3}$$

we see that $e = 1+\frac{u-k}{u-v}$.

Since $\boldsymbol{p}$ is a prefix of $\boldsymbol{v}$ and $z+p = k$, it follows that $\boldsymbol{k} = \boldsymbol{z}\boldsymbol{p}$. Thus we can also write $\boldsymbol{x} = (\boldsymbol{k}\boldsymbol{p}^{e-1})^2$.

Finally, setting $\boldsymbol{y} = \boldsymbol{u}[k+2v-u+1..u]$, since $\boldsymbol{z}\boldsymbol{y}$ is a suffix of $\boldsymbol{u}$ of length

$$k+v-u+2(u-v)-k = u-v = p,$$

it follows that $\boldsymbol{z}\boldsymbol{y}$ is a rotation of $\boldsymbol{p}$. In fact, $\boldsymbol{z}\boldsymbol{y} = \boldsymbol{q} = R_d(\boldsymbol{p})$, where by (3)

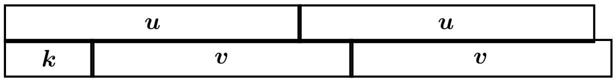
$$d = (u-z) \bmod (u-v) = (u-k) \bmod (u-v).$$

Then $\boldsymbol{z} = \boldsymbol{q}^f$, where

$$f = \frac{z}{p} = \frac{(k+v)-u}{u-v} = \frac{k}{u-v} - 1 \le 1,$$

so that $\boldsymbol{q}^f = \boldsymbol{q}[1..z]$ and $\boldsymbol{x} = (\boldsymbol{q}[1..k+v-u]\boldsymbol{p}^e)^2$, as required.

(c) Let $\boldsymbol{z} = \boldsymbol{u}[1..k+v-u]$, nonempty prefix of $\boldsymbol{u}$ and suffix of $\boldsymbol{v}$.



As in (b), observe that

$$\boldsymbol{u}[k+j] = \boldsymbol{v}[j] = \boldsymbol{u}[z+j],\ 1 \le j \le v-z = u-k,$$

with $z = k+v-u < k$. Again $\boldsymbol{u}[z+1..u]$ has period $k-z = u-v = p$, where $\boldsymbol{p} = \boldsymbol{u}[z+1..k] = \boldsymbol{v}[1..u-v]$. However, unlike (b), not $\boldsymbol{v}$, but only $\boldsymbol{v}[1..v-y]$, has period p, where $\boldsymbol{y} = \boldsymbol{v}[2u-(k+v)+1..v] = \boldsymbol{v}[u-z+1..v]$ and

$$y = k+2v-2u = z-(u-v) = z-p < z.$$

Thus, noting that $z < v$ and setting

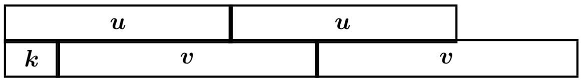
$$e = \frac{u-z}{u-v} = \frac{(u-v)+(u-k)}{u-v} = 1+\frac{u-k}{u-v} > 1,$$

we can write $\boldsymbol{x} = (\boldsymbol{z}\boldsymbol{p}^e)^2\boldsymbol{y}$. Since $\boldsymbol{u}[z+1..u]$ has prefix $\boldsymbol{p}$ and $k = z+p$, we see that $\boldsymbol{k} = \boldsymbol{z}\boldsymbol{p}$; further, since $\boldsymbol{z}$ has suffix $\boldsymbol{y}$ and $z = p+y$, it follows that $\boldsymbol{z} = \boldsymbol{q}\boldsymbol{y}$ for some rotation $\boldsymbol{q}$ of $\boldsymbol{p}$. In fact, $\boldsymbol{q} = R_d(\boldsymbol{p})$, where

$$d = (v-y) \bmod (u-v) = ((v-z)+(u-v)) \bmod (u-v) = (u-k) \bmod (u-v).$$

Noting that $\boldsymbol{k} = \boldsymbol{q}\boldsymbol{y}\boldsymbol{p}$, we see that both $\boldsymbol{k}\boldsymbol{v}$ and $\boldsymbol{x}$ have border $\boldsymbol{q}\boldsymbol{y}$, while $\boldsymbol{x} = (\boldsymbol{q}\boldsymbol{y}\boldsymbol{p}^e)^2\boldsymbol{y}$, as required.

(d) Again let $\boldsymbol{z} = \boldsymbol{u}[1..k+v-u]$, nonempty prefix of $\boldsymbol{u}$ and suffix of $\boldsymbol{u}$.



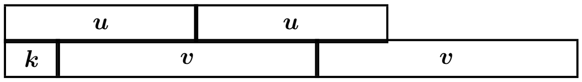
Observe that

$$\boldsymbol{u}[k+j] = \boldsymbol{v}[j] = \boldsymbol{u}[z+j],\ 1 \le j \le u-z,$$

where $z = k+v-u > k$. Thus $\boldsymbol{u}[k+1..u]$ and $\boldsymbol{v}[1..u-k]$ have period $z-k = v-u = p$. Therefore, setting $\boldsymbol{p} = \boldsymbol{v}[1..z-k]$, $e = \frac{u-k}{v-u}$, we can write $\boldsymbol{x} = (\boldsymbol{k}\boldsymbol{p}^e)^2\boldsymbol{y}$, where $\boldsymbol{y} = \boldsymbol{v}[2u-(k+v)+1..v]$ is a suffix of $\boldsymbol{v}$, $y = (k+v-u)+(v-u) = z+p > z$. Since $\boldsymbol{z}$ is a suffix of $\boldsymbol{y}$ and $y-z = p$, it follows that $\boldsymbol{y} = \boldsymbol{q}\boldsymbol{z}$, where $\boldsymbol{q} = R_d(\boldsymbol{p})$, $d = (u-k) \bmod (v-u)$. Similarly, since $\boldsymbol{k}$ is a prefix of $\boldsymbol{z}$ and $z-k = p$, we see that $\boldsymbol{z} = \boldsymbol{k}\boldsymbol{p}$, hence that $\boldsymbol{y} = \boldsymbol{q}\boldsymbol{k}\boldsymbol{p}$. Thus $\boldsymbol{x} = (\boldsymbol{k}\boldsymbol{p}^e)^2\boldsymbol{q}\boldsymbol{k}\boldsymbol{p}$, as claimed, and $\boldsymbol{x}$ and $\boldsymbol{k}\boldsymbol{v}$ both have border $\boldsymbol{k}\boldsymbol{p}$.

To see that $e > 1$, note that $k < \frac{v}{2}$, so that $e > \frac{v-u/2}{v-u} > 1$.

(e) Let $\boldsymbol{z} = \boldsymbol{v}[1..2u-(k+v)]$, nonempty prefix of $\boldsymbol{v}$ and suffix of $\boldsymbol{u}$.



Observe that

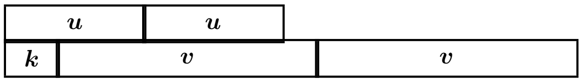
$$\boldsymbol{u}[k+j] = \boldsymbol{v}[j] = \boldsymbol{u}[(u-z)+j],\ 1 \le j \le z.$$

Thus $\boldsymbol{u}[k{+}1..u]$ has period $p = (u{-}z){-}k = v{-}u$. Consider

$$\boldsymbol{y} = \boldsymbol{v}[u{-}k{+}1..v] = \boldsymbol{u}[1..(k{+}v){-}u] = \boldsymbol{u}[1..k{+}p],$$

nonempty suffix of $\boldsymbol{v}$ and prefix of $\boldsymbol{u}$. Since $\boldsymbol{y}$ has prefix $\boldsymbol{k}$, it follows that $\boldsymbol{y} = \boldsymbol{kp}$, where $\boldsymbol{p} = \boldsymbol{u}[k{+}1..k{+}p]$. Thus for $e = \frac{u-k}{v-u}$, $\boldsymbol{v} = \boldsymbol{p}^e\boldsymbol{kp}$, $\boldsymbol{x} = \boldsymbol{k}(\boldsymbol{p}^e\boldsymbol{kp})^2$, as stated. Since $k{+}v < 2u$, $u{-}k > v{-}u$, and so $e > 1$.

(f) Let $\boldsymbol{z} = \boldsymbol{v}[2u{-}k{+}1..k{+}V]$, suffix of $\boldsymbol{v}$.



Observe that

$$\boldsymbol{v}[j] = \boldsymbol{u}[k{+}j] = \boldsymbol{v}[u{+}k{+}j],\ 1 \le j \le u{-}k,$$

so that $\boldsymbol{v}[1..2u{-}k]$ has period $p = u$. Then for $\boldsymbol{p} = \boldsymbol{u}[k{+}1..u]\boldsymbol{u}[1..k]$ and $e = \frac{2u-k}{u} \ge 1$, $\boldsymbol{v} = \boldsymbol{p}^e\boldsymbol{z}$ and $\boldsymbol{x} = \boldsymbol{k}(\boldsymbol{p}^e\boldsymbol{z})^2$. □

Case (f) of Lemma 2 is not a true overlap, but is included for completeness. Note also that cases (b), (c) and (e) require $k > 0$, and so do not exist if it is assumed that $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$ (or $\boldsymbol{v}^2$ and $\boldsymbol{w}^2$) occur at the same position.

We make the observation that if a third square $\boldsymbol{w}^2$ begins to the right of the starting position of $\boldsymbol{v}^2$, sufficiently near to satisfy the postulates of Lemma 2, then the analysis of the three squares $\boldsymbol{u}^2, \boldsymbol{v}^2, \boldsymbol{w}^2$ reduces to a simultaneous consideration of two of the lemma's cases. Thus, for example, the analysis of the situation shown in Figure 1 would take place in terms of the simultaneous occurrence of cases (d) (for $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$) and (b) (for $\boldsymbol{v}^2$ and $\boldsymbol{w}^2$). Indeed, all cases of three overlapping squares can be represented by pairs $[ij]$, $a \le i, j \le f$, referring to the cases (a)-(f) arising in Lemma 2. Figure 1 illustrates case $[db]$.

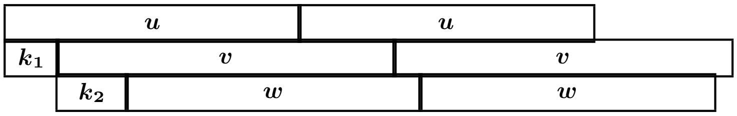


Fig. 1. $\boldsymbol{u^2}$ overlapping $\boldsymbol{v^2}$ (case (d)) that in turn overlaps $\boldsymbol{w^2}$ (case (b)): what is the combined effect?

Lemma 3 *In case $[db]$, if $u \ge 4(v{-}w)$, then $\boldsymbol{v}$ is a repetition of period $p = \gcd(v{-}u, v{-}w)$.*

Proof. We use subscripts $_1$ to identify variables for $\boldsymbol{u}$ and $\boldsymbol{v}$, subscripts $_2$ for those of $\boldsymbol{v}$ and $\boldsymbol{w}$. Observe then that for $e_1 > 1,\ e_2 \ge 1$,

$$\boldsymbol{x} = (\boldsymbol{k_1}\boldsymbol{p_1}^{e_1})^2\boldsymbol{q_1}\boldsymbol{k_1}\boldsymbol{p_1} = \boldsymbol{k_1}(\boldsymbol{q_2}[1..k_2{+}w{-}v]\boldsymbol{p_2}^{e_2})^2,$$

where the variables subscripted $_1$ relate to case (d) of Lemma 2, those subscripted $_2$ to case (b). It follows that $\boldsymbol{p_1}^{e_1}\boldsymbol{k_1}\boldsymbol{p_1}^{e_1}\boldsymbol{q_1}\boldsymbol{k_1}\boldsymbol{p_1} = \boldsymbol{v}^2$ must be a square, hence that $\boldsymbol{p_1}^{e_1}\boldsymbol{k_1}\boldsymbol{p_1} = \boldsymbol{p_1}^{e_1-1}\boldsymbol{q_1}\boldsymbol{k_1}\boldsymbol{p_1}$, where $\boldsymbol{q_1} = R_{d_1}(\boldsymbol{p_1})$. Applying the condition that

$$\boldsymbol{p_1}^{e_1}\boldsymbol{k_1}\boldsymbol{p_1} = \boldsymbol{q_2}[1..k_2{+}w{-}v]\boldsymbol{p_2}^{e_2},$$

and recalling that $\boldsymbol{z_1} = \boldsymbol{k_1}\boldsymbol{p_1}$ in case (d), while $z_2 = k_2{+}w{-}v$ in case (b), we find that a substring $\boldsymbol{v'}$ of $\boldsymbol{v}$ of length $v' = v{-}(z_1{+}z_2)$ has two periods, $p_1 = v{-}u$ and $p_2 = v{-}w$, where $\boldsymbol{v} = \boldsymbol{q_2}[1..z_2]\boldsymbol{v'}\boldsymbol{z_1}$. Thus, in order to apply the Periodicity Lemma [9], we compute

$$\begin{aligned} v{-}(z_1{+}z_2){-}(p_1{+}p_2) &= v{-}2v{+}u{+}w{-}z_1{-}z_2 \\ &= u{+}w{-}v{-}z_1{-}k_2{-}w{+}v \\ &= u{-}z_1{-}k_2. \end{aligned}$$

Since $z_1 = k_1{+}v{-}u$ and, from the definition of case (d) of Lemma 2, $k_1{+}v \le \frac{3u}{2}$, it follows that $z_1 \le \frac{u}{2}$; also, from Lemma 2(b), we conclude that $k_2 \le 2(v{-}w)$. Thus, $u{-}z_1{-}k_2 \ge \frac{u}{2}{-}2(v{-}w) \ge 0$ if

$$u \ge 4(v{-}w), \tag{4}$$

the condition that $\boldsymbol{v'}$ has period $p = \gcd(p_1, p_2)$. But then, observing that $\boldsymbol{v}$ has a prefix of period p_1 that includes $\boldsymbol{v'}$, as well as a suffix of period p_2 that includes $\boldsymbol{v'}$, we see that, if condition (4) holds, $\boldsymbol{v}$ itself must have period p. Since $\boldsymbol{v}$ has suffix $\boldsymbol{p_1}$, it must moreover be true that $\boldsymbol{v}$ is a repetition of period p, as required. □

The preceding lemma is a sample of the combinatorial information that may be obtained from considering all cases [ij] as specified above. To date, all the results given in [7, 8, 25, 17, 11] deal only with the special cases $[ij]$, $i = a, d$, that arise for $k_1 = 0$.

4 Commentary & Future Research

Obviously there is much work to be done to state and prove results such as Lemma 3 (which I believe can in fact be sharpened somewhat). In proving the results of [17], it turned out to be very helpful to look at the results of computer simulations for small values of k, u, v, w. It seems that similar techniques can profitably be used to generate conjectures for the cases $[ij]$ of three overlapping squares that arise from Lemma 2.

More generally, once the combinatorics of overlapping squares is well understood, it may well be possible to begin to design an algorithmic approach to the computation of runs that handles the various cases that arise without the need for elaborate preprocessing. This would for the first time permit direct analysis and computation of the local periodicities of a string in a manner consistent with their sparseness of occurrence.

References

1. Alberto Apostolico & Franco P. Preparata, **Optimal off-line detection of repetitions in a string**, *Theoret. Comput. Sci. 22* (1983) 297–315.
2. Gang Chen, Simon J. Puglisi & W. F. Smyth, **Fast & practical algorithms for computing all the runs in a string**, *Proc. 18th Annual Symp. Combinatorial Pattern Matching*, B. Ma & K. Zhang (eds.), LNCS 4580, Springer-Verlag (2007) 307–315.
3. Maxime Crochemore, **An optimal algorithm for computing all the repetitions in a word**, *Inform. Process. Lett. 12–5* (1981) 244–248.
4. Maxime Crochemore & Lucian Ilie, **Maximal repetitions in strings**, *J. Comput. Sys. Sci.* (2008) 796–807.
5. Maxime Crochemore, Lucian Ilie & Liviu Tinta, **Towards a solution to the "runs" conjecture**, *Proc. 19th Annual Symp. Combinatorial Pattern Matching*, P. Ferragina & G. Landau (eds.), LNCS 5029, Springer-Verlag (2008) 290–302.
6. Maxime Crochemore, Lucian Ilie & Liviu Tinta, **The "runs" conjecture**, *TCS 412–27* (2011) 2931–2941.
7. Maxime Crochemore and Wojciech Rytter, **Squares, cubes, and time-space efficient strings searching**, *Algorithmica 13* (1995), pp. 405–425.
8. Kangmin Fan, Simon J. Puglisi, W. F. Smyth & Andrew Turpin, **A new periodicity lemma**, *SIAM J. Discrete Math. 20–3* (2006) 656–668.
9. N. J. Fine and H. S. Wilf, **Uniqueness theorems for periodic functions**, *Proc. Amer. Math. Soc. 16* (1965) 109–114.
10. Aviezri S. Fraenkel & Jamie Simpson, **The exact number of squares in Fibonacci words**, *Theoret. Comput. Sci. 218–1* (1999) 95–106.
11. Frantisek Franek, Robert C. G. Fuller, Jamie Simpson & W. F. Smyth, **More results on overlapping squares**, *J. Discrete Algorithms* (2012) to appear.
12. Frantisek Franek, R. J. Simpson & W. F. Smyth, **The maximum number of runs in a string**, *Proc. 14th Australasian Workshop on Combinatorial Algs.*, Mirka Miller & Kunsoo Park (eds.) (2003) 26–35.
13. Mathieu Giraud, **Not so many runs in strings**, *Proc. 2nd Internat. Conf. on Language & Automata Theory & Applications*, Carlos Martín-Vide, Friedrich Otto & Henning Fernau (eds.), LNCS 5196, Springer-Verlag (2008) 232–239.
14. Mathieu Giraud, **Asymptotic behavior of the numbers of runs and microruns**, *Inform. & Computation* 207–11 (2009) 1221–1228.
15. Costas S. Iliopoulos & W. F. Smyth, **A characterization of the squares in a Fibonacci string**, *Theoret. Comput. Sci.* 172 (1997) 281–291.
16. Roman Kolpakov & Gregory Kucherov, **On maximal repetitions in words**, *J. Discrete Algorithms 1* (2000) 159–186.
17. Evguenia Kopylova & W. F. Smyth, **The three squares lemma revisited**, *J. Discrete Algorithms 11* (2012) 3–14.
18. Michael G. Main, **Detecting leftmost maximal periodicities**, *Discrete Applied Maths. 25* (1989) 145–153.
19. Michael G. Main & Richard J. Lorentz, **An $O(n \log n)$ algorithm for finding all repetitions in a string**, *J. Algorithms 5* (1984) 422–432.
20. Wataru Matsubara, Kazuhiko Kusano, Akira Ishino, Hideo Bannai & Ayumi Shinohara, **New lower bounds for the maximum number of runs in a string**, *PSC* (2008) 140–145.

21. Simon J. Puglisi & R. J. Simpson, **The expected number of runs in a word**, *Australasian J. Combinatorics 42* (2008) 45–54.
22. Simon J. Puglisi, R. J. Simpson & W. F. Smyth, **How many runs can a string contain?**, *Theoret. Comput. Sci. 401* (2008) 165–171.
23. Wojciech Rytter, **The number of runs in a string: improved analysis of the linear upper bound**, *Proc.* 23rd *Symp. Theoretical Aspects of Computer Science*, B. Durand & W. Thomas (eds.), LNCS 2884, Springer-Verlag (2006) 184–195.
24. SHARCNET, `https://www.sharcnet.ca/my/front/`
25. R. J. Simpson, **Intersecting periodic words**, *Theoret. Comput. Sci.* 374 (2007) 58–65.
26. Jamie Simpson, **Modified Padovan words and the maximum number of runs in a word**, *Australasian J. Combinatorics* 46 (2010) 129–145.
27. Bill Smyth, *Computing Patterns in Strings*, Pearson Addison-Wesley (2003) 423pp.