Theoretical Computer Science 596 (2015) 23–40

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Three overlapping squares: The general case characterized & applications

Widmer Bland [a], W.F. Smyth [a,b,*]

[a] Algorithms Research Group, Department of Computing & Software, McMaster University, Hamilton, Ontario L8S 4K1, Canada
[b] School of Engineering & Information Technology, Murdoch University, Murdoch WA 6150, Australia

ARTICLE INFO

Article history:
Received 13 March 2014
Received in revised form 15 June 2015
Accepted 18 June 2015
Available online 26 June 2015
Communicated by M. Crochemore

Keywords:
String
Word
Overlapping squares
Repetition
Run
Maximal periodicity

ABSTRACT

The "Three Squares Lemma" [9] 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 [10,30,21,13] have greatly extended this result to a "New Periodicity Lemma" (NPL) 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. We then apply these new insights, first to proofs of the final two remaining unproved subcases (out of a total of 14) of the NPL [10], then to an instance of the more general problem.

1. Introduction

Beginning with the "Three Squares Lemma" of Crochemore & Rytter [9], there has for several years been considerable interest in the limitations that may exist on periodicity in strings. An early survey of this topic by Mignosi & Restivo, with useful suggestions for future research directions, appears as Chapter 8 in [22]. In [9] it was shown 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. Over the last decade, a sequence of papers [10,30,21,13] 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 $k \geq 0$ to the right. First stated and proved as the "New Periodicity Lemma" (NPL) in [10], the main theorem has since been made more specific, with 12 of 14 subcases proved [30,21,13] — a main achievement of this paper is to establish the two that remain. Thus the assumption that three neighbouring squares of well-defined size exist within these well-defined bounds has been shown to lead to the conclusion that locally the string breaks down into repetitions of small period. In this paper we begin by proving a lemma that deals in a precise way with just two overlapping squares; we then apply this result to complete the proof of the final two cases of the NPL. We are as a consequence able to characterize the general case of three overlapping squares — no two constrained to begin at the same position — and therefore we can make a start on considering the combinatorial consequences.

[*] Corresponding author.
E-mail address: smyth@mcmaster.ca (W.F. Smyth).

http://dx.doi.org/10.1016/j.tcs.2015.06.037
0304-3975/© 2015 Elsevier B.V. All rights reserved.

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 [19] 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 has led to the conclusion that $\rho(n)$ is at least $0.9445757n$ [31,20] and no more than $n-1$ [2] — 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 — "probably nearby" because the interference of overlapping squares typically precludes periodic behaviour at one or more positions within the range of the double periodicity. 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 a form of brute force: global data structures (suffix array, longest common prefix array, Lempel–Ziv decomposition) need to be computed in an extended preprocessing phase, when of course runs are generally a *local* phenomenon. Moreover, it has been shown [26] that the *expected* number of runs in a string is much less than string length: runs generally occur sparsely. A global approach is necessitated by the absence of a detailed understanding of the combinatorics of overlapping occurrences of runs in strings.

In Section 2 terminology, notation and the relevant background are reviewed; Section 3 shows how to express the general case of three overlapping squares, making use of a careful analysis of two overlapping squares; Section 4 makes use of the new result to prove the two remaining subcases (3 & 7) of the NPL; then in Section 5 a further application to the general case of three overlapping squares is proved; finally, in Section 6 we briefly discuss future research directions.

2. Preliminaries

(Usage generally follows [32].) 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*** (or ***factor***) 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***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

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

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

If $\boldsymbol{v} = \boldsymbol{x}[i..j]$ has period p, where $v/p \geq 2$, and if neither $\boldsymbol{x}[i-1..j]$ nor $\boldsymbol{x}[i..j+1]$ (whenever these are defined) has period p, then the range $i..j$ in $\boldsymbol{x}$ is said to be a ***maximal periodicity*** or ***run*** in $\boldsymbol{x}$ [23]. A run is identified by a 4-tuple (i, p, e, t), where we choose p to be the ***minimum period*** of $\boldsymbol{v}$, $e = \lfloor v/p \rfloor \geq 2$ is its ***exponent***, and $t = v \bmod p \in 0..p-1$ is its ***tail***. Then $j = i + pe + (t-1)$. The string (1) has five runs

$$(1, 3, 2, 0),\ (1, 5, 2, 0),\ (3, 1, 2, 0),\ (4, 2, 2, 1),\ (8, 1, 2, 0)$$

corresponding to $(aba)^2, (abaab)^2, a^2, (ab)^2a, a^2$, respectively.

Every run in $\boldsymbol{x}$ determines $t+1$ ***repetitions***,

$$(i, p, e),\ (i+1, p, e), \ldots, (i+t, p, e),$$

where $(i', p, e),\ i \leq i' \leq i+t$, identifies the substring

$$\boldsymbol{x}[i'..i'+pe-1] = \boldsymbol{x}[i'..i'+p-1]^e.$$

Thus every repetition in $\boldsymbol{x}$ is a subrange of exactly one run in $\boldsymbol{x}$. For example, (1) has six repetitions

$$(1, 3, 2),\ (1, 5, 2),\ (3, 1, 2),\ (4, 2, 2),\ (5, 2, 2),\ (8, 1, 2)$$

corresponding to $(aba)^2, (abaab)^2, a^2, (ab)^2, (ba)^2, a^2$, respectively. Where no ambiguity arises, we will generally refer to runs and repetitions as substrings (for example, $(aba)^2$, $(ab)^2a$) rather than as ranges in $\boldsymbol{x}$ $(1..6, 4..8)$. If $e = 2$, we say that the repetition is a ***square***. We say that a square $\boldsymbol{u}^2$ is ***irreducible*** if $\boldsymbol{u}$ is not itself a repetition, ***regular*** if $\boldsymbol{u}$ has no square prefix.