Periods in Partial Words: An Algorithm

Francine Blanchet-Sadri\textsuperscript{1}, Travis Mandel\textsuperscript{2}, and Gautam Sisodia\textsuperscript{3}

\textsuperscript{1} Department of Computer Science, University of North Carolina, P.O. Box 26170, Greensboro, NC 27402–6170, USA blanchet@uncg.edu
\textsuperscript{2} Department of Mathematics, The University of Texas at Austin, 1 University Station, C1200, Austin, TX 78712, USA
\textsuperscript{3} Department of Mathematics, University of Washington, P.O. Box 354350, Seattle, WA 98195-4350, USA

Abstract. Partial words are finite sequences over a finite alphabet that may contain some holes. A variant of the celebrated Fine-Wilf theorem shows the existence of a bound $L = L(h, p, q)$ such that if a partial word of length at least $L$ with $h$ holes has periods $p$ and $q$, then it has period $\gcd(p, q)$. In this paper, we associate a graph with each $p$- and $q$-periodic word, and study two types of vertex connectivity on such a graph: modified degree connectivity and $r$-set connectivity where $r = q \mod p$. As a result, we give an algorithm for computing $L(h, p, q)$ in the general case.

1 Introduction

The problem of computing periods in words, or finite sequences of symbols from a finite alphabet, has important applications in several areas including data compression, coding, computational biology, string searching and pattern matching algorithms. Repeated patterns and related phenomena in words have played over the years a central role in the development of combinatorics on words [1], and have been highly valuable tools for the design and analysis of algorithms. In many practical applications, such as DNA sequence analysis, repetitions admit a certain variation between copies of the repeated pattern because of errors due to mutation, experiments, etc. Approximate repeated patterns, or repetitions where errors are allowed, are playing a central role in different variants of string searching and pattern matching problems [2]. Partial words, or finite sequences that may contain some holes, have acquired importance in this context. A (strong) period of a partial word $u$ over an alphabet $A$ is a positive integer $p$ such that $u(i) = u(j)$ whenever $u(i), u(j) \in A$ and $i \equiv j \mod p$ (in such a case, we call $u$ $p$-periodic). In other words, $p$ is a period of $u$ if for all positions $i$ and $j$ congruent modulo $p$, the letters in these positions are the same or at least one of these positions is a hole.

\textsuperscript{*} This material is based upon work supported by the National Science Foundation under Grant No. DMS–0452020.
There are many fundamental results on periods of words. Among them is the well-known periodicity result of Fine and Wilf [3], which determines how long a \( p \) - and \( q \)-periodic word needs to be in order to also be \( \gcd(p, q) \)-periodic. More precisely, any word having two periods \( p, q \) and length at least \( p + q - \gcd(p, q) \) has also \( \gcd(p, q) \) as a period. Moreover, the length \( p + q - \gcd(p, q) \) is optimal since counterexamples can be provided for shorter lengths, that is, there exists an optimal word of length \( p + q - \gcd(p, q) - 1 \) having \( p \) and \( q \) as periods but not having \( \gcd(p, q) \) as period [1]. Extensions of Fine and Wilf’s result to more than two periods have been given. For instance, in [4], Constantinescu and Ilie give an extension for an arbitrary number of periods and prove that their lengths are optimal.

Fine and Wilf’s result has been generalized to partial words [5,7,8,9,10,11]. Some of these papers are concerned with weak periodicity, a notion not discussed in this paper. The papers that are concerned with strong periodicity refer to the basic fact, proved by Shur and Konovalova (Gamzova) in [10], that for positive integers \( h, p \) and \( q \), there exists a positive integer \( l \) such that a partial word \( u \) with \( h \) holes, two periods \( p \) and \( q \), and length at least \( l \) has period \( \gcd(p, q) \). The smallest such integer is called the optimal length and it will be denoted by \( L(h, p, q) \). They gave a closed formula for the case where \( h = 2 \) (the cases \( h = 0 \) or \( h = 1 \) are implied by the results in [3,5]), while in [9], they gave a formula in the case where \( p = 2 \) as well as an optimal asymptotic bound for \( L(h, p, q) \) in the case where \( h \) is “large.” In [7], Blanchet-Sadri et al. gave closed formulas for the optimal lengths when \( q \) is “large,” whose proofs are based on connectivity in the so-called \( (p, q) \)-periodic graphs. In this paper, we study two types of vertex connectivity in these graphs: the modified degree connectivity and \( r \)-set connectivity where \( r = q \mod p \). Although the graph-theoretical approach is not completely new, the paper gives insights into periodicity in partial words and provides an algorithm for determining \( L(h, p, q) \) in all cases.

We end this section by reviewing basic concepts on partial words. Fixing a nonempty finite set of letters or an alphabet \( A \), finite sequences of letters from \( A \) are called (full) words over \( A \). The number of letters in a word \( u \), or length of \( u \), is denoted by \( |u| \). The unique word of length 0, denoted by \( \varepsilon \), is called the empty word. The set of all words over \( A \) of finite length is denoted by \( A^* \). A partial word \( u \) of length \( n \) over \( A \) is a partial function \( u : \{0, \ldots, n - 1\} \to A \). For \( 0 \leq i < n \), if \( u(i) \) is defined, then \( i \) belongs to the domain of \( u \), denoted by \( i \in D(u) \), otherwise \( i \) belongs to the set of holes of \( u \), denoted by \( i \in H(u) \). For convenience, we will refer to a partial word over \( A \) as a word over the enlarged alphabet \( A_\circ = A \cup \{\circ\} \), where \( \circ \notin A \) represents a “do not know” symbol or hole.

2 \((p, q)\)-Periodic Graphs

In this section, we discuss the fundamental property of periodicity, the goal of our paper which is to describe an algorithm to compute \( L(h, p, q) \) in all cases, and some initial results. We can restrict our attention to the case where \( p \) and \( q \) are coprime, since it is well-known that the general case can be reduced to the coprime case (see, for example, [5,9]). Also, we assume without loss of generality