

Binary De Bruijn Partial Words with One Hole^{*}

F. Blanchet-Sadri¹, J. Schwartz², S. Stich³, and B. J. Wyatt¹

¹ Department of Computer Science, University of North Carolina,
P.O. Box 26170, Greensboro, NC 27402–6170, USA, blanchet@uncg.edu

² Department of Computer Science, Princeton University,
35 Olden Street, Princeton, NJ 08540–5233, USA

³ Department of Mathematics, Princeton University,
Fine Hall, Washington Road, Princeton, NJ 08544–1000, USA

Abstract. In this paper, we investigate partial words, or finite sequences that may have some undefined positions called holes, of maximum subword complexity. The subword complexity function of a partial word w over a given alphabet of size k assigns to each positive integer n , the number $p_w(n)$ of distinct full words over the alphabet that are *compatible* with factors of length n of w . For positive integers n, h and k , we introduce the concept of a de Bruijn partial word of order n with h holes over an alphabet A of size k , as being a partial word w with h holes over A of minimal length with the property that $p_w(n) = k^n$. We are concerned with the following three questions: (1) What is the length of k -ary de Bruijn partial words of order n with h holes? (2) What is an efficient method for generating such partial words? (3) How many such partial words are there? *Keywords:* Combinatorics on words; Partial words; Subword complexity; De Bruijn sequences; De Bruijn graphs; Eulerian paths; combinatorial problems.

1 Introduction

Let A be a k -letter alphabet and w be a finite or right infinite word over A . A *subword* or *factor* of w is a block of consecutive letters of w . The subword complexity of w is the function that assigns to each positive integer, n , the number, $p_w(n)$, of distinct subwords of length n of w . The subword complexity, also called symbolic complexity, of finite and infinite words has become an important subject in combinatorics on words. Application areas include dynamical systems, ergodic theory, and theoretical computer science. We refer the reader to Chapter 10 of [1] which surveys and discusses subword complexity of finite and infinite words. References [2] and [3] provide other surveys, [4] shows how the so-called special and bispecial factors can be used to compute the subword complexity, and [5] gives another interesting approach based on the gap function.

* This material is based upon work supported by the National Science Foundation under Grant No. DMS–0754154. The Department of Defense is also gratefully acknowledged. We thank the referees of preliminary versions of this paper for their very valuable comments and suggestions.

When we restrict our attention to finite words of maximum subword complexity, de Bruijn sequences play an important role. A k -ary de Bruijn sequence of order n is a word over an alphabet of size k where each of the k^n words of length n over the alphabet appears as a factor exactly once. It is well known that such sequences have length $k^n + n - 1$. There are $k!^{k^{n-1}}$ of them, and they can be efficiently generated by constructing Eulerian cycles in corresponding de Bruijn directed graphs. The technical report of de Bruijn provides an history on the existence of these sequences [6]. De Bruijn graphs find applications, in particular, in genome rearrangements [7], in the complexity of deciding avoidability of sets of partial words [8], etc.

In this paper, we investigate partial words of maximum subword complexity. Partial words are finite sequences that may have some undefined positions called holes (a (full) word is just a partial word without holes). Partial words can be viewed as sequences over an extended alphabet $A_\diamond = A \cup \{\diamond\}$, where $\diamond \notin A$ stands for a hole. Here \diamond matches every letter in the alphabet, or is compatible with every letter in the alphabet. For example, $10\diamond 01$ is a partial word with one hole over the alphabet $\{0, 1\}$. In this context, $p_w(n)$ is the number of distinct full words over the alphabet that are compatible with factors of length n of the partial word w (in our example with $w = 10\diamond 01$, we have $p_w(3) = 5$ since $000, 001, 010, 100$ and 101 match factors of length 3 of w). For positive integers n, h and k , we introduce the concept of a de Bruijn partial word of order n with h holes over an alphabet A of size k , as being a partial word w with h holes over A of minimal length with the property that $p_w(n) = k^n$.

The contents of our paper is as follows: In Section 2, we review some concepts on partial words. In Section 3, we give lower and upper bounds on the length of k -ary de Bruijn partial words with h holes of order n , and show that our bounds are tight when $h = 1$. In Section 4, we provide an algorithm to construct de Bruijn binary partial words with one hole. Finally in Section 5, we show how to count such partial words by adapting the so-called BEST theorem that counts the number of Eulerian cycles in directed graphs.

2 Preliminaries

For more background on partial words, we refer the reader to [9].

Let A be a fixed non-empty finite set called an *alphabet* whose elements we call *letters*. A *word* over A is a finite sequence of elements from A . We let A^* denote the set of words over A which, under the concatenation operation of words, forms a free monoid whose identity is the empty word, which we denote by ε . Unless otherwise stated, we assume that A contains at least two letters.

A *partial word* w of length n over A can be defined as a function $w : [0..n - 1] \rightarrow A_\diamond$, where $A_\diamond = A \cup \{\diamond\}$ with $\diamond \notin A$. The length of w is denoted by $|w|$, and $w(i)$, the symbol at position i , is denoted by w_i (here $[0..n - 1]$ denotes the set of positions $\{0, 1, \dots, n - 1\}$). For $0 \leq i < n$, if $w(i) \in A$, then i belongs to the *domain* of w , denoted $D(w)$, and if $w(i) = \diamond$, then i belongs to the *set of holes* of w , denoted $H(w)$. Whenever $H(w)$ is empty, w is a *full word*. We refer

to an occurrence of the symbol \diamond as a *hole*. We let A_\diamond^* denote the set of all partial words over A .

A partial word u is a *factor* of the partial word v if there exist x, y such that $v = xuy$. The factor u is called *proper* if $u \neq \varepsilon$ and $u \neq v$. The partial word u is a *prefix* (respectively, *suffix*) of v if $x = \varepsilon$ (respectively, $y = \varepsilon$).

The partial word u is *contained* in the partial word v , denoted $u \subset v$, if $|u| = |v|$ and $u(i) = v(i)$ for all $i \in D(u)$. Two partial words u and v of equal length are *compatible*, denoted $u \uparrow v$, if $u(i) = v(i)$ whenever $i \in D(u) \cap D(v)$. In other words, u and v are compatible if there exists a partial word w such that $u \subset w$ and $v \subset w$, in which case we let $u \vee v$ denote the *least upper bound* of u and v ($u \subset (u \vee v)$ and $v \subset (u \vee v)$ and $D(u \vee v) = D(u) \cup D(v)$). For example, $u = aba\diamond$ and $v = a\diamond b\diamond$ are compatible, and $(u \vee v) = abab\diamond$.

A full word u is a *subword* of w if there exists some $0 \leq i < |w| - |u|$ such that $u \uparrow w(i) \cdots w(i + |u| - 1)$. Informally, under some “filling in” of the holes in w with letters from A to form the full word w' , there is some consecutive block of letters in w' , $w'(i) \cdots w'(i + |u| - 1)$, such that $w'(i) = u(0), w'(i + 1) = u(1)$, and so on. Note that in this paper, subwords are always full.

A completion \hat{w} of a partial word w over A is a function $\hat{w} : [0..|w| - 1] \rightarrow A$ such that $\hat{w}(i) = w(i)$ if $w(i) \neq \diamond$. A completion \hat{w} is usually thought of as a “filling in” of the holes of w with letters from A . Note that two partial words u and v are compatible if there exist completions \hat{u} and \hat{v} such that $\hat{u} = \hat{v}$. The subword complexity of w is the function that assigns to each integer, $0 \leq n \leq |w|$, the number, $p_w(n)$, of distinct full words over A that are compatible with factors of length n of w (or the number of distinct subwords of w of length n). We let $\text{Sub}_w(n)$ denote the set of all subwords of w of length n , and we let $\text{Sub}(w) = \bigcup_{0 \leq n \leq |w|} \text{Sub}_w(n)$ the set of all subwords of w . Note that if \hat{w} is a completion of w , then $p_{\hat{w}}(n) \leq p_w(n)$, since $\text{Sub}_{\hat{w}}(n) \subset \text{Sub}_w(n)$.

3 Bounds on the length of de Bruijn partial words

What is the length of a shortest word w over an alphabet of size k for which $p_w(n) = k^n$, where n is a positive integer?

Theorem 1 ([1]). *For all $k, n \geq 1$ there exists a word w over an alphabet of size k , of length $k^n + n - 1$, such that $p_w(n) = k^n$.*

Such a word is often called a *k-ary de Bruijn full word of order n*, that is, a full word over a given alphabet A with size k for which every possible word of length n over A appears as a subword exactly once. De Bruijn words are often “cyclic” in the literature, meaning that subwords can wrap around from the end to the beginning of the word, but to better fit our notion of the complexity function, we unwrap them and use a non-cyclic version.

In order to prove the theorem, set $A = \{0, 1, \dots, k - 1\}$. If $k = 1$, then take 0^n , while if $n = 1$, take $01 \cdots (k - 1)$. If $k, n \geq 2$, a family of directed graphs $G_k(n)$ is defined as follows: the vertices of $G_k(n)$ are the words of length $n - 1$ over A , and the edges of $G_k(n)$ are the pairs (az, zb) , labelled by azb , where $a, b \in A$ and

z is a word of length $n - 2$ over A . It then suffices to show that $G_k(n)$ possesses an Eulerian cycle, that is, a path that traverses every edge exactly once and begins and ends at the same vertex. Indeed, $G_k(n)$ is strongly connected, that is, there is a directed path connecting any two vertices, and the indegree of each vertex equals its outdegree. A directed graph that possesses an Eulerian cycle is called an Eulerian digraph. Note that there are several linear-time algorithms, including Fleury's algorithm, for computing Eulerian cycles in digraphs [10].

We define a k -ary de Bruijn partial word with h holes of order n , to be a partial word of minimal length with h holes over an alphabet A of size k with all k^n words of length n over A being compatible with factors of it. A main question is to determine the length of k -ary de Bruijn partial words with h holes of order n . For example, 00110 is a 2-ary de Bruijn full word of order 2, which has length 5, while a 2-ary de Bruijn partial word of order 2 with one hole is 001 \diamond , which has length 4. We let $L_k(n, h)$ denote the length of a k -ary de Bruijn partial word of order n with h holes.

Definition 1. – Let $M_z(n)$ denote the number of distinct completions of factors of length n with at least one hole of a partial word z .
– Let $M_k(n, h) = \max_z M_z(n)$ where the maximum is taken over all partial words z with h holes over an alphabet of size k .

It is clear that for $n \leq h$, if z is a word with h holes, n of them being consecutive, over an alphabet of size k , then $M_z(n) = k^n$ and since k^n is the total number of words of length n over a k -letter alphabet, we have $M_k(n, h) = k^n$. We can significantly refine the upper bound of k^n on $M_k(n, h)$, when $n > h$, as stated in the next theorem.

Theorem 2. For $k \geq 2$ and $n > h > 0$, $M_k(n, h) \leq (n - h + 1)k^h + 2\frac{k^h - k}{k - 1}$.

Proof. Let z be a word with h holes over a k -letter alphabet. First, note that if the h holes of z are consecutive, or \diamond^h is a factor of z , then there may be factors of length n of z that contain only the first hole (respectively, the last hole), the first two holes (respectively, the last two holes), and so on, until may be factors that contain only the first $h - 1$ holes (respectively, the last $h - 1$ holes), and then factors that contain all of the h holes. Note that i consecutive holes can contribute a maximum of k^i distinct completions. So, in total, z can have up to $(n - h + 1)k^h + 2\sum_{i=1}^{h-1} k^i = (n - h + 1)k^h + 2\frac{k^h - k}{k - 1}$ distinct completions of factors of length n containing at least one hole. Now, assume that \diamond^{h-r} and \diamond^r are two disjoint factors of z , where $0 < r < h$. In this case, $M_z(n)$ cannot be bigger than the bound above. So if we keep splitting up the holes, we do not change our bound. \square

Corollary 1. For $k \geq 2$, $n \geq 2h + 2$ and $h > 0$, we have

$$M_k(n, h) = (n - h + 1)k^h + 2\frac{k^h - k}{k - 1}.$$

Proof. By Theorem 2, $M_k(n, h) \leq (n-h+1)k^h + 2\frac{k^h-k}{k-1}$. To show that $M_k(n, h) \geq (n-h+1)k^h + 2\frac{k^h-k}{k-1}$, we only need find a partial word $z_{n,h}$ with h holes over a k -letter alphabet such that $M_{z_{n,h}}(n) = (n-h+1)k^h + 2\frac{k^h-k}{k-1}$. Consider $z_{n,h} = b^n a \diamond^h ab^n$ where a, b are distinct letters of the alphabet. The factors of length n of $z_{n,h}$ with at least one hole are

- $b^{n-2}a\diamond, \dots, b^{n-h}a\diamond^{h-1}$, as well as their reversals: the number of distinct completions of these factors is $2\frac{k^h-k}{k-1}$.
- $b^{n-h-1}a\diamond^h, \dots, ba\diamond^h ab^{n-h-3}, a\diamond^h ab^{n-h-2}, \diamond^h ab^{n-h-1}$: the number of distinct completions of these factors is $(n-h-1+1+1)k^h = (n-h+1)k^h$.

Note that the words of length n compatible with these factors are distinct, since the factors starting at the first $n-1$ positions are distinct from each other, because they start with a different number of b 's, and distinct from the rest, because they have an a at most h positions from the beginning. The factors ending at the last $n-1$ positions are also distinct because they end with different numbers of b 's. \square

Remark 1. Corollary 1 fails for $n < 2h + 2$. In the construction of the proof of Corollary 1, $z_{5,2} = b^5 a \diamond^2 ab^5$, and so $M_z(5) = 19 \neq 20 = (n-h+1)k^h + 2\frac{k^h-k}{k-1}$. Here, the factor $b^2 a \diamond^2$ is compatible with the factor $\diamond^2 ab^2$, and so the completion $b^2 ab^2$ gets counted twice.

Theorem 3. For $h > 0$, $L_k(n, h) \geq L_k(n, 0) - M_k(n, h) + (n + h - 1)$.

Proof. A k -ary de Bruijn full word of order n contains each subword of length n exactly once. When considering partial words with h holes over an alphabet of size k , we are still limited to at most one distinct factor of length n per starting symbol, except we can get more than one distinct completion for factors with at least one hole. The number of such completions is at most $M_k(n, h)$, but this includes $(n+h-1)$ starting positions that lead to distinct subwords in a de Bruijn full word. So, in total we have $L_k(n, h) \geq L_k(n, 0) - M_k(n, h) + (n + h - 1)$. \square

Corollary 2. For $n \geq 2h+2$ and $h > 0$, $L_2(n, h) \geq 2^n + 2n + h + 2 - (n - h + 3)2^h$.

Proof. We know that 2-ary de Bruijn full words of order n have length $2^n + n - 1$. Furthermore, from Theorem 2, Corollary 1 and Theorem 3, we get

$$\begin{aligned} L_2(n, h) &\geq 2^n + n - 1 - (2^h(n - h + 3) - 4) + (n + h - 1) \\ &= 2^n + 2n + h + 2 - (n - h + 3)2^h \end{aligned} \quad \square$$

In Section 4, we show that the bound of Corollary 2 is tight for $h = 1$, that is, $L_2(n, 1) = 2^n + 2n + h + 2 - (n - h + 3)2^h = 2^n - 1$ for $n \geq 4$.

4 Constructing de Bruijn partial words

We can construct k -ary de Bruijn full words of order n by finding Eulerian cycles in $G_k(n)$. In this section, we describe an algorithm to construct 2-ary de Bruijn

partial words of order n with one hole by finding Eulerian paths in a trimmed version of $G_2(n)$. We also discuss the $k = 3$ case.

We first recall the conditions for a directed graph $G = (V, E)$ to have an (x, y) -Eulerian path, that is, an Eulerian path from vertex x to vertex y . Let $\text{iddeg}(v)$ (respectively, $\text{odeg}(v)$) denote the indegree (respectively, outdegree) of vertex $v \in V$.

Lemma 1. *Let $G = (V, E)$ be a digraph, and let $x, y \in V$ be such that $\text{odeg}(x) = 1 + \text{iddeg}(x)$ and $\text{iddeg}(y) = 1 + \text{odeg}(y)$. Then G has an (x, y) -Eulerian path if and only if G has at most one non-trivial connected component containing x, y and for every vertex $v \in V \setminus \{x, y\}$, $\text{iddeg}(v) = \text{odeg}(v)$.*

We now modify the Eulerian cycle approach to prove that our bounds are tight in the binary one hole case.

Theorem 4. *For $n \geq 4$, we have $L_2(n, 1) = 2^n - 1$.*

Proof. Start with the digraph $G = G_2(n)$. Let $z = x \diamond y = 1^{n-2}0 \diamond 0^{n-2}1$. It can be checked that $M_z(n) = 2n = M_2(n, 1)$, that is, z has $2n$ distinct subwords of length n . Trim $G_2(n)$ by deleting all edges that are in $\text{Sub}_z(n)$. Then, add a new edge from vertex x to vertex y labelled by z . Call the resulting graph, $G' = (V, E)$. First, consider any factor of length $n - 1$ with a hole in z . Then, choose a completion, v , of that factor. Thus, v is a prefix of some $v_1 \in \text{Sub}_z(n)$ and a suffix of some $v_2 \in \text{Sub}_z(n)$. So, both $\text{iddeg}(v)$ and $\text{odeg}(v)$ get decreased by one, but v remains balanced. The only vertices that become isolated are 0^{n-1} and 10^{n-2} . Now, consider the factors $x = 1^{n-2}0$ and $y = 0^{n-2}1$. Here, x is a prefix of two subwords of length n , namely the two completions $1^{n-2}00$ and $1^{n-2}01$. So, two edges starting at x are deleted from G , while the edge starting at x , labelled by z , is added to G . Similarly, y is a suffix of two subwords of length n , the two completions $00^{n-2}1$ and $10^{n-2}1$. So, two edges ending at y are deleted from G , while the edge ending at y , labelled by z , is added to G .

So the graph G' satisfies the following conditions: (1) G' has a single non-trivial connected component; (2) $\text{odeg}(y) = 1 + \text{iddeg}(y)$ and $\text{iddeg}(x) = 1 + \text{odeg}(x)$; and (3) for every vertex $v \in V \setminus \{x, y\}$, $\text{iddeg}(v) = \text{odeg}(v)$. By Lemma 1, G' has an Eulerian path from vertex y to vertex x . Since z has the maximum number, $M_2(n, 1)$, of distinct subwords of length n , we get $L_2(n, 0) = 2^n + n - 1$ implies $L_2(n, 1) = 2^n - M_2(n, 1) + n - 1 + (|y| + 1)$, and so $L_2(n, 1) = 2^n - 2n + n - 1 + n = 2^n - 1$ as desired. \square

Example 1. Computer experiments show that there are seven z 's (up to a renaming of letters) with one hole over the binary alphabet $\{0, 1\}$ such that $M_z(4) = M_2(4, 1) = 8$. They are $110 \diamond 110, 110 \diamond 011, 110 \diamond 001, 101 \diamond 001, 100 \diamond 101, 100 \diamond 100$ and $100 \diamond 011$. From the proof of Theorem 4, if we choose $z_1 = 110 \diamond 001$, then there is an Eulerian path in the resulting graph from vertex 001 to vertex 110 . But, if we consider $z_2 = 110 \diamond 110$ instead, we note that $1110, 0111 \in \text{Sub}(z_2)$ but $1111 \notin \text{Sub}(z_2)$. So the vertex 111 becomes isolated with the loop labelled by 1111 (see the graph on the right in Figure 1). Therefore, the resulting graph

does not have an Eulerian path. There are fourteen z 's that satisfy $M_z(4) = 8$, but only four of them generate graphs that have an Eulerian path (110◊001, its reversal, and their renamings).

What we need is to start with a word z that is *good* in the sense that if $\{10^{n-1}, 0^{n-1}1\} \subset \text{Sub}_z(n)$ then $0^n \in \text{Sub}_z(n)$, and if $\{01^{n-1}, 1^{n-1}0\} \subset \text{Sub}_z(n)$ then $1^n \in \text{Sub}_z(n)$, otherwise 0^{n-1} or 1^{n-1} would become isolated with loop 0^n or 1^n , respectively. Table 1 gives data on the number of z 's over the alphabet $\{0, 1\}$ such that $M_z(n) = 2n$ versus the number of such z 's that are good.

Table 1. Number of good z 's over $\{0, 1\}$ for $4 \leq n \leq 8$

n	Number of z 's such that $M_z(n) = 2n$	Number of good z 's
4	14	4
5	98	10
6	546	40
7	2768	96
8	12832	272

After applying the algorithm described in the proof of Theorem 4 (see Algorithm 1), we get a 2-ary de Bruijn partial word of order n of length $2^n - 1$ with one hole. We let $G_2(n, z)$ denote the graph built by Algorithm 1.

Algorithm 1 Constructing a 2-ary de Bruijn word of order n with one hole, where $n \geq 4$

- 1: Build $G = G_2(n)$
 - 2: Select a good word $z = x \diamond y$ with $|x| = |y| = n - 1$ and $M_z(n) = 2n$
 - 3: Compute $S = \text{Sub}_z(n)$
 - 4: Create graph G' from G by deleting the edges in the set S along with any resulting isolated vertices, and add an edge from vertex x to vertex y labelled by z
 - 5: Find an Eulerian path p in G' from y to x
 - 6: **return** p
-

Example 2. For $k = 2$ and $n = 4$, if we select $z_1 = 110 \diamond 001$ then Algorithm 1 produces the graph on the left in Figure 1. The 2-ary word $w = 0010110 \diamond 0011110$ of length $2^4 - 1 = 15$ is such that $p_w(4) = 2^4 = 16$. It can be checked that

$$\begin{array}{cccccccc}
 001 & \xrightarrow{0010} & 010 & \xrightarrow{0101} & 101 & \xrightarrow{1011} & 011 & \xrightarrow{0110} & 110 & \xrightarrow{110 \diamond 001} & 001 \\
 & & \xrightarrow{0011} & & \xrightarrow{0111} & & \xrightarrow{1111} & & \xrightarrow{1110} & & \\
 & & & & 011 & & 111 & & 111 & & 110
 \end{array}$$

is an Eulerian path from $y = 001$ to $x = 110$ in the trimmed graph $G_2(4, z_1)$.

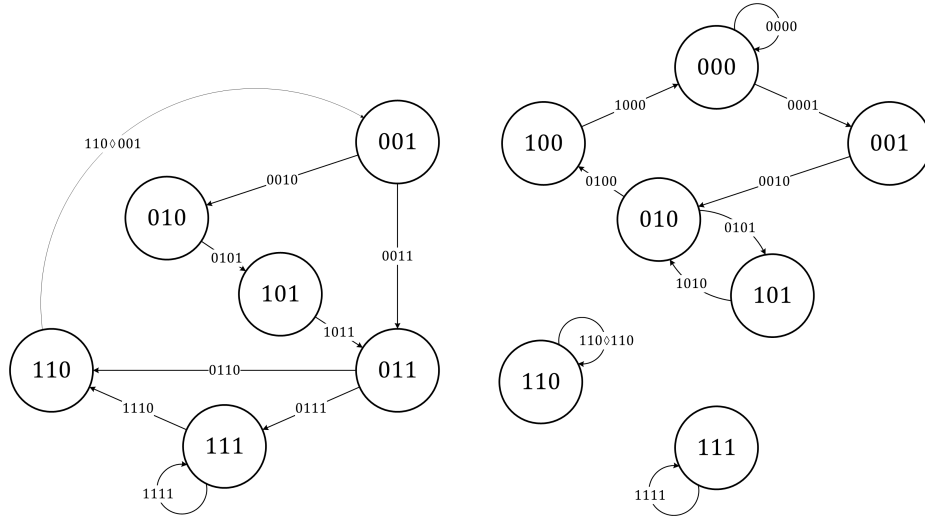


Fig. 1. Left: Non-trivial connected component of $G_2(4, 110 \diamond 001)$; Right: Non-trivial connected components of $G_2(4, 110 \diamond 110)$

The difficulty in building a de Bruijn partial word for $k = 3$, $n = 4$, and $h = 1$ for instance, is that we have to compensate for the indegree and outdegree of the nodes connected by the edge labelled by a word with one hole. When working with a good word $z = x \diamond y$, thus having the maximum number $kn = 3n$ of subwords of length n , a “schism” is created in which $\text{ideg}(x) = 3$ and $\text{odeg}(x) = 1$, while $\text{ideg}(y) = 1$ and $\text{odeg}(y) = 3$. For example, if we take $z = 110 \diamond 001$, the node 110 will now have outdegree 1 because of the edge labelled by z , and indegree 3, while 001 only has indegree 1 but will have outdegree 3. With $k = 2$, we compensate for this schism by starting the de Bruijn partial word from y (that has outdegree 2 and indegree 1), and by ending with x (that has indegree 2 and outdegree 1). Since each vertex, other than x and y are balanced, this effectively “skips” the problem entirely. When we try to compensate in a similar fashion for a 3-letter alphabet, we end up having to add an extra edge.

To produce a de Bruijn partial word in which a single subword occurs twice, use a good word of the form $x \diamond x$. For example, using $102 \diamond 102$ eliminates all edges to and away from the node 102. This removes the issue of compensating for an unbalanced vertex: each vertex has equal indegree and outdegree (note that $\text{ideg}(102) = \text{odeg}(102) = 1$ due to the edge $102 \diamond 102$ from 102 to 102). However, the vertex 102 becomes isolated. Since all edges from 102 have been deleted, an additional edge is required to connect 102 to the rest of the graph. Here, use the edge $(102, 022)$ labelled by 1022 for instance. This process can be generalized to arbitrary n , and so we get the following result.

Theorem 5. For $n \geq 2$, we have $L_3(n, 1) = 3^n - n$.

Proof. The equality $L_3(n, 0) = 3^n + n - 1$ implies $L_3(n, 1) = (3^n - M_3(n, 1) + |z|) + 1$ (for an extra edge), and so $L_3(n, 1) = 3^n - 3n + 2n - 1 + 1 = 3^n - n$. \square

Example 3. The partial word $u = 10\circ 1002221222020121112000$ is a 3-ary de Bruijn partial word of length 24 with one hole of order $n = 3$, while

$$v = 102\circ 10222212220220122112200212120210121 \\ 11121002020012011200001110110010100022$$

of length 77 is one of order $n = 4$. Note that u has the subword 100 occurring twice, while v has the subword 1022 occurring twice as explained above.

5 Counting de Bruijn partial words

Another main question is to compute the number of k -ary de Bruijn partial words with h holes of order n , which we denote by $N_k(n, h)$. It is well known that $N_k(n, 0) = k!^{k^{n-1}}$, which can be calculated by counting the number of Eulerian cycles in $G_k(n)$. This can be done by using the so-called BEST theorem, named after de Bruijn, van Aardenne-Ehrenfest, Smith and Tutte, that counts the number of Eulerian cycles in directed graphs.

Theorem 6 ([11]). *Let $G = (V, E)$ be an Eulerian digraph, and let L_G denote the Laplacian matrix of G defined as follows: for $i = j$, $L_G(i, j) = \text{odeg}(v_i) - e$, and for $i \neq j$, $L_G(i, j) = -e$, where e is the number of edges from v_i to v_j . Then the number of non-equivalent Eulerian cycles in G is*

$$C \prod_{v \in V} (\text{odeg}(v) - 1)! = C \prod_{v \in V} (\text{iddeg}(v) - 1)! \tag{1}$$

with C any cofactor of L_G .

To compute $N_2(n, 1)$, we need to modify Theorem 6, since we want to count the number of Eulerian paths.

Theorem 7. *Let $G = (V, E)$ be a digraph, and let $x, y \in V$ be such that $\text{odeg}(x) = 1 + \text{iddeg}(x)$ and $\text{iddeg}(y) = 1 + \text{odeg}(y)$. Suppose that G satisfies the conditions of Lemma 1 to have an (x, y) -Eulerian path. Let L_G denote the Laplacian matrix of G defined as above. Then the number of (x, y) -Eulerian paths in G is given by (1) with C the cofactor of L_G with the row and column corresponding to vertex y removed.*

With 2-ary de Bruijn partial words of order n with one hole, as mentioned in Section 4, we need to apply Theorem 6 to more than one graph since every word z of length $2n - 1$, with a hole in the middle and such that $M_z(n) = M_2(n, 1) = 2n$, can potentially serve as the new edge added to the graph $G_2(n)$. But after deleting the edges corresponding to subwords of length n of z , we do not necessarily have an Eulerian path, so we must only count those paths in the $G_2(n, z)$'s, where z is good. This suggests an algorithm, Algorithm 2, to count the number of 2-ary de Bruijn partial words of order n with one hole.

Algorithm 2 Computing the number $N_2(n, 1)$, where $n \geq 4$

- 1: Find the set Z of all good z 's of the form $x \circ y$ such that $|x| = |y| = n - 1$ and $M_z(n) = M_2(n, 1) = 2n$
 - 2: **for all** $z \in Z$ **do**
 - 3: Construct the Laplacian matrix $L_z = L_{G_2(n, z)}$
 - 4: Eliminate all rows and columns of L_z that have all zero entries
 - 5: Calculate the determinant of the matrix L_z after removing the row and column that correspond to x
 - 6: **return** The sum of the determinants
-

Remark 2. Step 4 is necessary since some vertices may have become isolated. This still would allow for Eulerian paths, but would make the determinant zero if those rows and columns were left in the Laplacian matrix. We also eliminate the row and column corresponding to x to form the cofactor, since by Theorem 4, x must be the last vertex of the path because $\text{iddeg}(x) = \text{oddeg}(x) + 1$. In step 5, the $(\text{iddeg}(x) - 1)!$ multiplicative factor is always 1 since $\text{iddeg}(x) = 2$. Unfortunately, unlike the full case where the sum falls out easily since all cofactors of the single matrix have the same value, the cofactors of the L_z 's may be different.

Example 4. Returning to Example 1 with $k = 2$ and $n = 4$, up to reversal and renaming of letters, we only need to consider $z_1 = 110 \circ 001$ to compute $N_2(n, 1)$. Referring to the graph on the left in Figure 1, $L_{G_2(4, z_1)}$ is as follows:

	001	010	011	101	110	111
001	2	-1	-1	0	0	0
010	0	1	0	-1	0	0
011	0	0	2	0	-1	-1
101	0	0	-1	1	0	0
110	-1	0	0	0	1	0
111	0	0	0	0	-1	1

Note that the rows and columns corresponding to the vertices 000 and 100 have been removed since all their entries are zeros. If we remove the row and column of vertex 110, we get a determinant of 4. So there are 4 Eulerian paths from 001 to 110 in $G_2(4, z_1)$: $00110 \circ 001011110$, $0011110 \circ 00101110$, $00101110 \circ 0011110$ and $001011110 \circ 001110$. Since the only z 's that are good are $110 \circ 001$, its reversal, and their renamings, we get $N_2(4, 1) = 4 \times 4 = 16$.

References

1. Allouche, J.P., Shallit, J.: Automatic Sequences: Theory, Applications, Generalizations. Cambridge University Press (2003)
2. Allouche, J.P.: Sur la complexité des suites infinies. Bulletin of the Belgian Mathematical Society **1** (1994) 133–143
3. Ferenczi, S.: Complexity of sequences and dynamical systems. Discrete Mathematics **206** (1999) 145–154

4. Cassaigne, J.: Complexité et facteurs spéciaux. *Bulletin of the Belgium Mathematical Society* **4** (1997) 67–88
5. Gheorghiciuc, I.: The subword complexity of a class of infinite binary words. *Advances in Applied Mathematics* **39** (2007) 237–259
6. De Bruijn, N.G.: Acknowledgement of priority to C. Flye Sainte-Marie on the counting of circular arrangements of $2n$ zeros and ones that show each n -letter word exactly once. Technical Report 75–WSK–06, Department of Mathematics and Computing Science, Eindhoven University of Technology, The Netherlands (1975)
7. Alekseyev, M.A., Pevzner, P.A.: Colored de Bruijn graphs and the genome halving problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **4**(1) (2007) 98–107
8. Blakeley, B., Blanchet-Sadri, F., Gunter, J., Rampersad, N.: On the complexity of deciding avoidability of sets of partial words. In Diekert, V., Nowotka, D., eds.: *DLT 2009, 13th International Conference on Developments in Language Theory*, Stuttgart, Germany. Volume 5583 of *Lecture Notes in Computer Science.*, Berlin, Heidelberg, Springer-Verlag (2009) 113–124 www.uncg.edu/cmp/research/unavoidablesets3.
9. Blanchet-Sadri, F.: *Algorithmic Combinatorics on Partial Words*. Chapman & Hall/CRC Press, Boca Raton, FL (2008)
10. Gross, J.L., Yellen, J.: *Handbook of Graph Theory*. CRC Press (2004)
11. Stanley, R.P.: *Enumerative Combinatorics, Vol. 2. Volume 2*. Cambridge University Press, Cambridge (2001)