WYATT, BENJAMIN J., M.S. De Bruijn Partial Words. (2013)
Directed by Dr. Francine Blanchet-Sadri. 48 pp.

In a k^n-complex word over an alphabet Σ of size k each of the k^n words of length n appear as a subword at least once. Such a word is said to have maximum subword complexity. De Bruijn sequences of order n over Σ are the shortest words of maximum subword complexity and are well known to have length k^n+n-1. They are efficiently constructed by finding Eulerian cycles in so-called de Bruijn graphs.

In this thesis, we investigate partial words, or sequences with wildcard symbols or hole symbols, 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 h, k and n, a de Bruijn partial word of order n with h holes over an alphabet Σ of size k is a partial word w with h holes over Σ of minimal length with the property that $p_w(n) = k^n$. In some cases, they are efficiently constructed by finding Eulerian paths in modified de Bruijn graphs. 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?

DE BRUIJN PARTIAL WORDS

by

Benjamin J. Wyatt

A Thesis Submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Greensboro
2013

Approved by

_____
Committee Chair

APPROVAL PAGE

This thesis has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____
Francine Blanchet-Sadri

Committee Members _____
Fereidoon Sadri

_____
Jing Deng

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

Foremost, I would like to express the deepest appreciation to my committee chair, Dr. Francine Blanchet-Sadri, for her continuous support of my research, for her patience, enthusiasm, and motivation. Her wisdom and guidance helped me in all of my time of research and study, and I could not have worked with a better advisor and mentor for my Master's Degree study.

I would like to also thank the members of my thesis committee: Dr. Fereidoon Sadri and Dr. Jing Deng.

My sincere graditude goes to Dr. Steve Tate, Dr. Nancy Green, Dr. Shan Suthaharan, Dr. Lixin Fu, Mark Armstrong, and Lydia Fritz for providing guidance, direction, and opportunities that encouraged me to pursue this work.

I wish to thank the co-authors of the two publications which form the foundation of this thesis, Derek Allums, John Lensmire, Jarett Schwartz, and Slater Stich, for their research and collaboration.

And last, but not least, I thank Jeffrey Collis for providing both inspiration and unconditional support, and for just being there for me.

This material is based upon work supported by the National Science Foundation under Grant Nos. DMS-0754154 and DMS-1060775. The Department of Defense is also gratefully acknowledged.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Let Σ be a k-letter alphabet and w be a finite or right infinite word over Σ. 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 readers to Chapter 10 of [3] which surveys and discusses subword complexity of finite and infinite words. References [2] and [11] provide other surveys, [9] shows how the so-called special and bispecial factors can be used to compute the subword complexity, and [12] gives another interesting approach based on the gap function.

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 generated in linear time 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 [10]. De Bruijn graphs find applications, in particular, in genome rearrangements [1], in the complexity of deciding avoidability of sets of partial words [4], etc.

In this thesis, 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 $\Sigma_\diamond = \Sigma \cup \{\diamond\}$, where $\diamond \notin \Sigma$ stands for a hole. Here $\diamond$ matches every letter in the alphabet, or is compatible with every letter in the alphabet. For example, $10\diamond01$ 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 $w = 10\diamond01$, 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 Σ of size k, as being a partial word w with h holes over Σ of minimal length with the property that $p_w(n) = k^n$.

The contents of this thesis is as follows. In Chapter I, we start by introducing partial words, along with de Bruijn full words and digraphs. In Chapter II, we discuss de Bruijn partial words with one hole, starting with such words over a binary alphabet, then a ternary alphabet, then the general case over a k-ary alphabet, followed by a discussion of counting binary de Bruijn partial words with one hole. In Chapter III we look at de Bruijn partial words with two or more holes by first discussing minimal partial words of maximal subword complexity, then by looking at closed formulas for known lengths of de Bruijn partial words. In Chapter IV, we consider results from a case where the length of a de Bruijn partial word is unknown, followed by open questions. Part of this thesis has been published in [6] and in [8].

1.1 Preliminaries

For more background on partial words, we refer readers to [5].

Let Σ be a fixed non-empty finite set called an *alphabet* whose elements we call *letters*. A *word* over Σ is a finite sequence of elements from Σ. We let Σ^* denote the set of words over Σ 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 Σ contains at least two letters.

A *partial word* w of length n over Σ can be defined as a function $w : [0..n-1] \to \Sigma_\diamond$, where $\Sigma_\diamond = \Sigma \cup \{\diamond\}$ with $\diamond \notin \Sigma$. 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 $\{0, 1, \ldots, n-1\}$). For $0 \leq i < n$, if $w(i) \in \Sigma$, 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 $\Sigma_\diamond^*$ denote the set of all partial words over Σ.

A partial word u is a *factor* of the partial word w if there exist x, y such that $w = xuy$. The factor u is called *proper* if $u \neq \varepsilon$ and $u \neq w$. The partial word u is a *prefix* (respectively, *suffix*) of w if $x = \varepsilon$ (respectively, $y = \varepsilon$). For $i = 0, \ldots, h$, let $F_i(w, n)$ denote the multiset containing the factors of w of length n with exactly i holes.

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, $u = aba\diamond\diamond$ and $v = a\diamond\diamond b\diamond$ 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 Σ 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 thesis, subwords are always full.

A completion $\hat{w}$ of a partial word w over Σ is a function $\hat{w} : [0..|w| - 1] \to \Sigma$ 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 Σ. 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 Σ that are compatible with factors of length n of w (or the number of distinct subwords of w of length n). We let $\mathrm{Sub}_w(n)$ denote the set of all subwords of w of length n, and we let $\mathrm{Sub}(w) = \bigcup_{0 \leq n \leq |w|} \mathrm{Sub}_w(n)$ 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 $\mathrm{Sub}_{\hat{w}}(n) \subset \mathrm{Sub}_w(n)$. For $|\Sigma| = k$, if all subwords of length n are subwords of w, then $\mathrm{Sub}_w(n) = k^n$ and we call w a *k^n-complex word*. Alternatively, we say that w has maximum subword complexity.

We end this section by recalling a bound on the subword complexity of any binary partial word w. Let $n \leq |w|$ be a positive integer. A factor u of length n of w is repeated, if there exist integers $i \neq j$ such that $u = w(i) \cdots w(i + n - 1) = w(j) \cdots w(j + n - 1)$. Similarly, a subword u of length n of w is repeated, if there exist integers $i \neq j$ such that $u \uparrow w(i) \cdots w(i + n - 1)$ and $u \uparrow w(j) \cdots w(j + n - 1)$.

Note that repeated factors imply repeated subwords, but the converse does not hold in general.

Lemma I.1 ([7]). *Let w be a partial word with h holes over a binary alphabet. For index $i = 0, \ldots, h$ and positive integer $n \leq |w|$, let $F_i(w, n)$ denote the multiset containing the factors of w of length n with exactly i holes. Then*

$$\sum_{i=0}^{h} \|F_i(w, n)\| = |w| - n + 1 \tag{I.1}$$

$$p_w(n) \leq \sum_{i=0}^{h} 2^i \|F_i(w, n)\| \tag{I.2}$$

with equality holding in (I.2) if and only if w has no repeated subwords of length n. The following zero-hole and one-hole bounds hold:

(1) *Let $h = 0$. For $n \leq |w|$, we have $p_w(n) \leq |w| - n + 1$, with equality holding if and only if w has no repeated subwords of length n.*

(2) *Let $h = 1$ and $n \leq |w|$. If $|w| \leq 2n - 1$, then $p_w(n) \leq 2(|w| - n + 1)$. Else, $p_w(n) \leq |w| + 1$. In both cases, equality holds if and only if w has no repeated subwords of length n.*

1.2 De Bruijn 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 I.2 ([3]). *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 Σ with size k for which every possible word of length n over Σ 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 $\Sigma = \{0, 1, \ldots, k-1\}$. 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 Σ, and the edges of $G_k(n)$ are the pairs (az, zb), labelled by azb, where $a, b \in \Sigma$ and z is a word of length $n-2$ over Σ. 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, as each Eulerian cycle in $G_k(n)$ corresponds to a k-ary de Bruijn full word of order n. The graph $G_k(n)$ has an Eulerian cycle because it is both *weakly-connected* and *balanced*. A digraph G is weakly-connected if, after replacing each edge in G with a non-directional edge between the same vertices, the resulting graph is connected. A vertex v is *balanced* if its indegree, denoted $\texttt{ideg}(v)$, equals its outdegree, $\texttt{odeg}(v)$ (we call the pair $(\texttt{ideg}(v), \texttt{odeg}(v))$ the degree of v, or $\texttt{deg}(v)$). If every vertex v in G is balanced, then G is a balanced digraph.

Indeed, $G_k(n)$ is strongly connected, that is, there is a directed path connecting any two vertices, and $G_k(n)$ is balanced. 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 [13].

Example I.3. The 2-ary word $w = 0000111101100101000$ of length 19 is such that $p_w(4) = 2^4 = 16$. It can be checked from Figure 1 that

$$000 \xrightarrow{0000} 000 \xrightarrow{0001} 001 \xrightarrow{0011} 011 \xrightarrow{0111} 111 \xrightarrow{1111} 111 \xrightarrow{1110} 110 \xrightarrow{1101} 101 \xrightarrow{1011} 011$$
$$\xrightarrow{0110} 110 \xrightarrow{1100} 100 \xrightarrow{1001} 001 \xrightarrow{0010} 010 \xrightarrow{0101} 101 \xrightarrow{1010} 010 \xrightarrow{0100} 100 \xrightarrow{1000} 000$$

is an Eulerian cycle in $G_2(4)$.

It is well known that there are $k!^{k^{n-1}}$ de Bruijn full words of order n over a k-letter alphabet. Returning to Example I.3, this gives 256 de Bruijn full words of order 4 over $\{0, 1\}$.

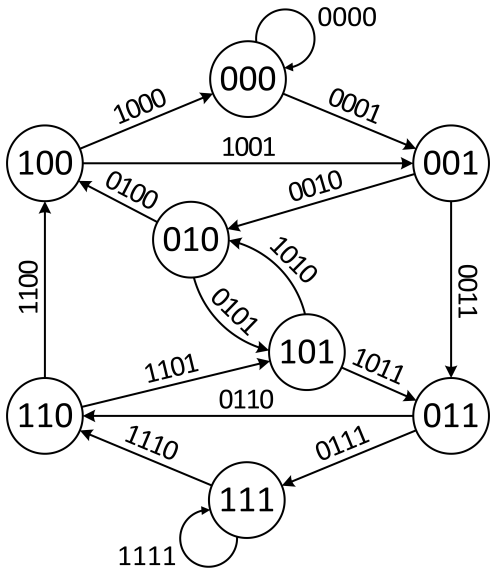

Figure 1. $G_2(4)$, a de Bruijn full word graph

A de Bruijn partial word of order n with h holes over Σ of size k is a k^n-complex word with h holes over Σ of shortest possible length. 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 I.4 ([8]). We define $M_z(n)$ and $M_k(n,h) = \max_z M_z(n)$ as follows.

- 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 I.5 ([8]). *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. □

Corollary I.6 ([8]). *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 I.5, $M_k(n,h) \leq (n-h+1)k^h + 2\frac{k^h-k}{k-1}$. 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

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

Remark. Corollary I.6 fails for $n < 2h+2$. In the construction of the proof of Corollary I.6, $z_{5,2} = b^5a\diamond^2ab^5$, and so $M_z(5) = 19 \neq 20 = (n-h+1)k^h + 2\frac{k^h-k}{k-1}$. Here, the factor $b^2a\diamond^2$ is compatible with the factor $\diamond^2ab^2$, and so the completion b^2ab^2 gets counted twice.

Theorem I.7 ([8]). *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 $L_k(n,h) \geq L_k(n,0) - M_k(n,h) + (n+h-1)$. □

Corollary I.8 ([8]). *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 I.5, Corollary I.6 and Theorem I.7, we get

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

□

In Chapter II, we show that the bound of Corollary I.8 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 $n \geq 4$.

CHAPTER II

DE BRUIJN PARTIAL WORDS WITH ONE HOLE

We can construct a k-ary de Bruijn full word of order n by finding an Eulerian cycle in the de Bruijn graph $G_k(n)$. To construct a k-ary de Bruijn partial word of order n with h holes, we show that we can delete edges from $G_k(n)$ and add edges to $G_k(n)$ in such a way that a minimal length Eulerian path exists. In this chapter we discuss in detail our technique as applied to the special cases of binary and ternary de Bruijn partial words with one hole, as well as k-ary de Bruijn partial words with one hole in the general case; each of order n.

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. Eulerian paths are found in weakly-connected and "nearly" balanced, or *weakly-balanced*, digraphs. Let $G = (V, E)$ be a digraph with vertices u and v such that $\texttt{ideg}(u) = \texttt{odeg}(u) + 1$ and $\texttt{ideg}(v) + 1 = \texttt{odeg}(v)$. If every other vertex in V is balanced, then G is weakly-balanced.

Lemma II.1 ([13]). *Let G be a digraph with vertices u and v, where $\texttt{ideg}(u) = \texttt{odeg}(u) + 1$ and $\texttt{ideg}(v) + 1 = \texttt{odeg}(v)$. Then G has an Eulerian path from v to u if and only if (1) G is weakly-connected and (2) G is weakly-balanced.*

Before we look at constructing such a digraph, we consider the placement of holes, which is a key factor in building a de Bruijn partial word. One effective strategy for determining hole placement is to "encapsulate" the holes in a partial word $z = xz'y$, where z meets the following conditions:

- factors x and y are full words;
- $|x| = |y| = n - 1$;
- the first and last positions of z' contain a hole; and
- z' contains h holes.

Then for each partial word w with h holes containing z as a factor, we have $M_z(n) = M_w(n)$, as this means that $F_i(w, n) = F_i(z, n)$ for $i \geq 1$. We call each such partial word z a *candidate word*, and we call each candidate word that occurs in a de Bruijn partial word of matching parameters a *good word*.

In Section 2.1, we describe an algorithm to construct a binary de Bruijn partial word of order n with one hole by finding an Eulerian path in a trimmed version of $G_2(n)$. We then discuss the $k = 3$ case in Section 2.2 and the general case for $k > 3$ in Section 2.3.

2.1 The binary one-hole case

We now modify the Eulerian cycle approach to prove that our lower bound is tight in the binary one hole case.

Theorem II.2 ([8]). *For $n \geq 4$, we have $L_2(n, 1) = 2^n - 1$.*

Proof. Start with the digraph $G = G_2(n)$. Let $z = x \diamond y = 1^{n-2}0 \diamond 0^{n-2}1$. Trim $G_2(n)$ by deleting all edges that are in $\mathrm{Sub}_z(n)$. Then, add a new edge from vertex x to vertex y labelled by z. Call the resulting graph, $G' = (V, E)$, the digraph generated by z. We show that there exists a path from any vertex u in G' to $x = 1^{n-2}0$. To see this, if $u = u_0 \cdots u_{n-3}1$, we traverse the path labelled by the

edges $u_0 \cdots u_{n-3}11, u_1 \cdots u_{n-3}111, \ldots, u_{n-3}1^{n-2}0$, which starts at vertex u and ends at vertex x. None of the edges in this path are deleted from G as can be seen from the form of z. Similarly, if $u = u_0 \cdots u_{n-3}0$, we traverse the path labelled by the edges

$$u_0 \cdots u_{n-3}01, u_1 \cdots u_{n-3}011, \ldots, u_{n-3}01^{n-2}, 01^{n-2}0$$

which starts at u and ends at x. There are only three problem vertices u, that is, $1^{n-2}0, 10^{n-2}$ and 0^{n-1}, since indeed both out-edges are deleted for each of these and we thus cannot follow the edge labelled $u1$ to get to x. However, $1^{n-2}0$ is our x and is accounted for. Additionally, 10^{n-2} and 0^{n-1} become isolated and are thus removed from G. Thus, G' has a single non-trivial connected component.

It can be checked that $M_z(n) = 2n = M_2(n, 1)$, that is, z has $2n$ distinct subwords of length n. 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 $v_1 \in \mathrm{Sub}_z(n)$ and a suffix of some $v_2 \in \mathrm{Sub}_z(n)$. So, both $\texttt{ideg}(v)$ and $\texttt{odeg}(v)$ get decreased by one, but v remains balanced. As already mentioned, 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 (see Figure 2). 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) $\texttt{odeg}(y) = \texttt{ideg}(y) + 1$ and $\texttt{ideg}(x) = \texttt{odeg}(x) + 1$; and

(3) for every vertex $v \in V \setminus \{x, y\}$, $\texttt{ideg}(v) = \texttt{odeg}(v)$.

Then conditions (2) and (3) imply that G' is weakly-balanced. Thus, by Lemma II.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. □

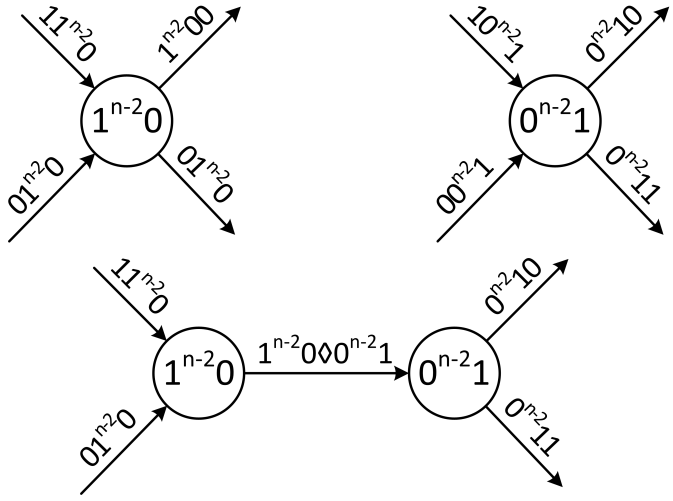

Figure 2. Proof of Theorem II.2.

Example II.3. 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

$$001\diamond001, 001\diamond100, 001\diamond110, 010\diamond110, 011\diamond010, 011\diamond011, 011\diamond100$$

From the proof of Theorem II.2, if we choose $z_1 = 001\diamond110$, then there is an Eulerian path in the resulting graph from vertex 110 to vertex 001. But, if we consider $z_2 = 001\diamond001$ instead, we note that $0001, 1000 \in \mathrm{Sub}(z_2)$ but $0000 \notin \mathrm{Sub}(z_2)$. So the vertex 000 becomes isolated with the loop labelled by 0000 (see the graph on the right in Figure 6). 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 ($001\diamond110$, its reversal, and their renamings). For $k = 2, n = 5, h = 1$, there are ninety eight z's that satisfy $M_z(5) = M_2(5,1) = 10$, but only fifty of them generate graphs that have an Eulerian path.

From this we can define what makes a candidate word z "good" in the binary one-hole case: $M_z(n) = M_2(n,1)$ and $G_2(n,z)$, the graph constructed from $G_2(n)$ after deleting and adding edges according to z (see the construction of $G' = G_2(n,z)$ in the proof of Theorem II.2), must be connected. Three classes of candidate words have been identified as problem words and computer experiments show that up to $n = 10$ these are indeed the only candidate words which are not good words in the binary one-hole case:

- $\{01^{n-1}, 1^{n-1}0\} \subset \mathrm{Sub}_z(n)$ and $1^n \notin \mathrm{Sub}_z(n)$, or $\{10^{n-1}, 0^{n-1}1\} \subset \mathrm{Sub}_z(n)$ and $0^n \notin \mathrm{Sub}_z(n)$. If one of these occurs, then 0^{n-1} or 1^{n-1} becomes isolated with loop 0^n or 1^n, respectively (see Figure 3).

- $z = x \diamond x$, in which case if we generate $G_2(n, z)$, all two in-edges to vertex x and two out-edges from x are deleted but as prescribed by Algorithm 1, an edge labelled $x \diamond x$ is added from x to x, resulting in a non-trivial disconnected component (see Figure 4).

- $z = 1^{n-2}0 \diamond 01^{n-2} = x \diamond y$, up to a renaming of letters, in which case if we build $G_2(n, z)$, the edge from y to x remains, both in-edges of y are deleted, both out-edges of x are deleted, an edge from x to y is added, and all four edges of 1^{n-1} remain (with one out-edge connected to x and one in-edge connected to y), resulting in a disconnected component with three vertices and five edges (see Figure 5).

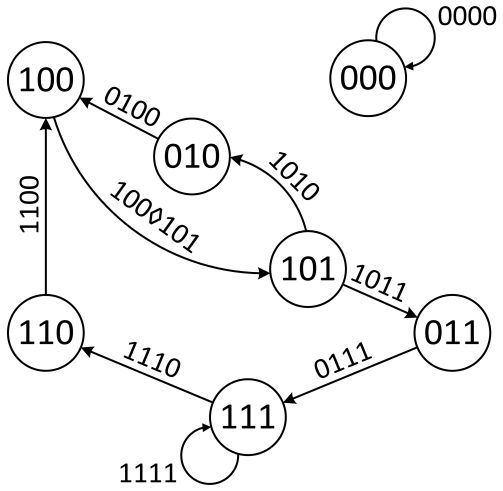

Figure 3. First class of problem words

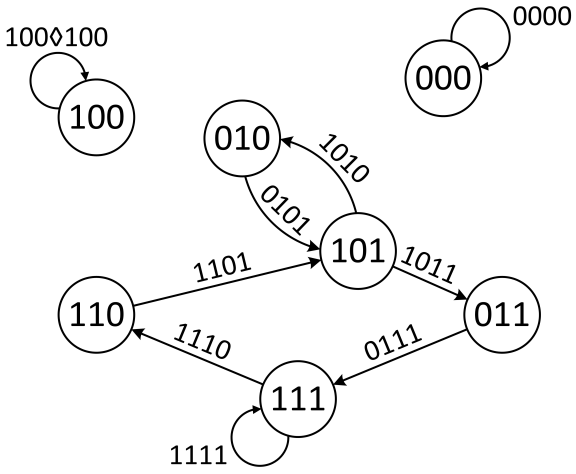

Figure 4. Second class of problem words

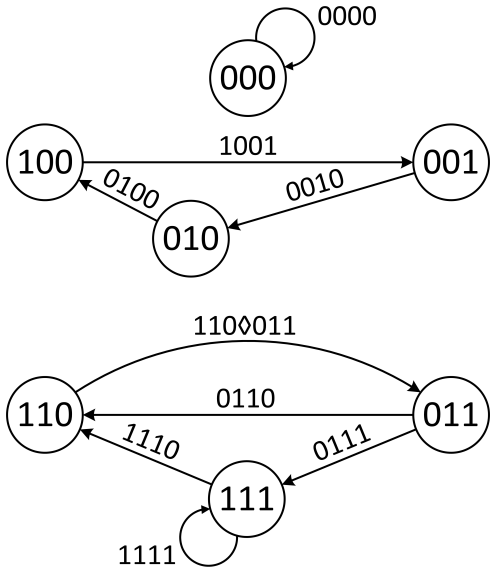

Figure 5. Third class of problem words

This brings us to a conjecture that these three classes of problem words are the only ones which induce disconnected components of $G_2(n, z)$.

Conjecture II.4. *Let $n \geq 4$, and let $z = x \diamond y$ such that $M_z(n) = 2n$. Then z is either a problem word of the first, second, or third class, or z is a good word.*

Table 1 gives data on the number of candidate words z over the alphabet $\{0, 1\}$ such that $M_z(n) = 2n$ versus the number of such candidate words that are good words.

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	50
6	546	434
7	2768	2444
8	12832	12098
9	56352	54550
10	239088	235072

After applying the algorithm described in the proof of Theorem II.2 (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. Note that Algorithm 1 is efficient, as an Eulerian path can be found in linear time.

Algorithm 1 ([8]) 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 $S = \mathrm{Sub}_z(n)$
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 II.5 (H)**.** For $k = 2$ and $n = 4$, if we select $z_1 = 001\diamond110$ then Algorithm 1 produces the graph on the left in Figure 6.

The 2-ary word $w = 1101001\diamond1100001$ of length $2^4 - 1 = 15$ is such that $p_w(4) = 2^4 = 16$. It can be checked that

$$110 \xrightarrow{1101} 101 \xrightarrow{1010} 010 \xrightarrow{0100} 100 \xrightarrow{1001} 001 \xrightarrow{001\diamond110} 110$$
$$\xrightarrow{1100} 100 \xrightarrow{1000} 000 \xrightarrow{0000} 000 \xrightarrow{0001} 001$$

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

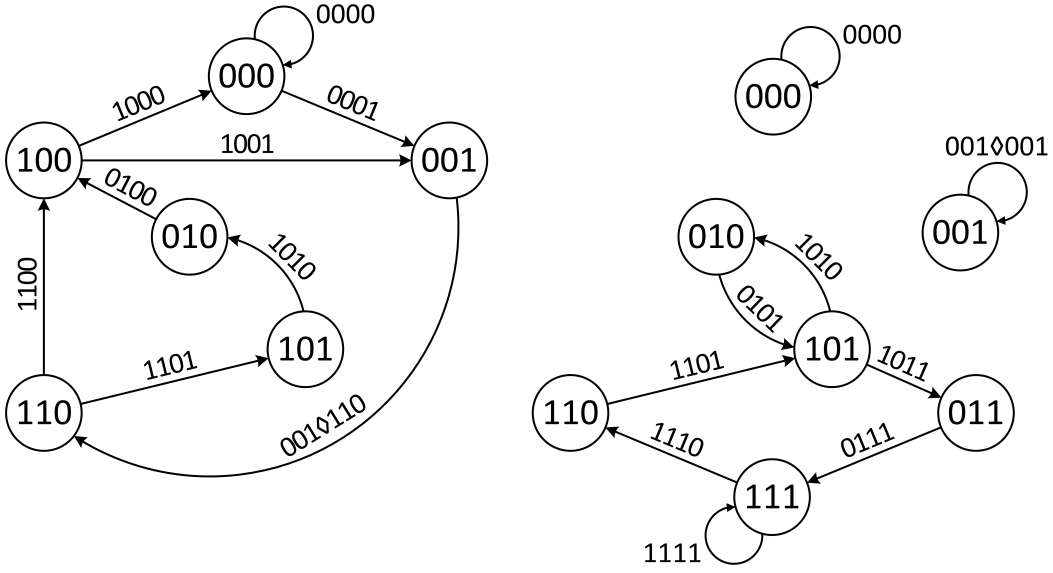

Figure 6. Left: $G_2(4, 001\diamond110)$; Right: $G_2(4, 001\diamond001)$

2.2 The ternary one-hole case

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. If we select a candidate word $z = x \diamond y$ that is a good word according to the criteria for the binary one-hole case, thus having the maximum number $kn = 3n$ of subwords of length n, a "schism" is created in which $\texttt{ideg}(x) = 3$ and $\texttt{odeg}(x) = 1$, while $\texttt{ideg}(y) = 1$ and $\texttt{odeg}(y) = 3$. For example, if we take $z = 001 \diamond 110$, the node 001 will now have outdegree 1 because of the edge labelled by z, and indegree 3, while 110 only has indegree 1 but will have outdegree 3 (see Figure 7).

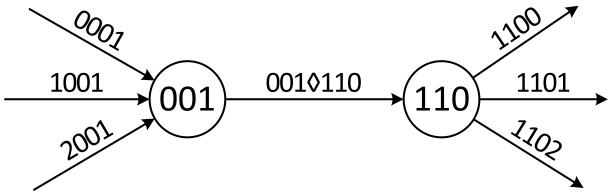

Figure 7. Vertices 001 and 110 in $G_3(4, 001 \diamond 110)$

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 is 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 $012 \diamond 012$ eliminates all edges to and away from the node 012. This removes the issue of compensating for an unbal-

anced vertex: each vertex has equal indegree and outdegree (note that $\texttt{ideg}(012) = \texttt{odeg}(012) = 1$ due to the edge 012◇012 from 012 to 012). However, the vertex 012 becomes isolated. Since all edges from 012 have been deleted, an additional edge is required to connect 012 to the rest of the graph. Here, use the edge $(012, 122)$ labelled by 0122 for instance (see Figure 8). This process can be generalized to arbitrary n, as described in Theorem II.6.

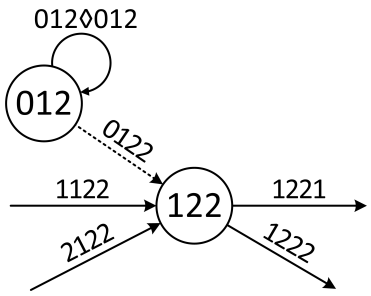

Figure 8. Vertex 012 from $G_3(4, 012\diamond012)$ with added edge 0122

Theorem II.6 ([8]). *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 $L_3(n, 1) = 3^n - 3n + 2n - 1 + 1 = 3^n - n$. □

Example II.7. The partial word $u = 01\diamond0102000222022111221210$ is a 3-ary de Bruijn partial word of length 24 with one hole of order $n = 3$ (see Figure 11), while

$$\begin{aligned} v \;=\; & 012\diamond012222022212210220022112020212010200 \\ & 0002011212110210021111000100110101112 2 \end{aligned}$$

of length 77 is one of order $n = 4$. Note that u has the subword 010 occurring twice, while v has the subword 0122 occurring twice as explained above.

For a good word of the form $z = x\diamond x$ in the ternary one-hole case, we have $M_z(n) = M_3(n,1) = M_k(n,1)$; this characteristic is shared with good words in the binary one-hole case. However, some good words of the form $z = x\diamond y$ exist in the ternary one-hole case, in which one subword occurs twice in z (see Example II.8). For such a good word $z = x\diamond y$ with $M_z(n) = M_3(n,1) - 1$, we can construct a de Bruijn partial word graph without adding an extra edge. Thus we have two algorithms for constructing a 3-ary de Bruijn partial word of order n with h holes, both with linear complexity. Algorithm 2 is used when a good word of the form $z = x\diamond x$ is selected.

Algorithm 2 Constructing a 3-ary de Bruijn word of order $n \geq 2$ with one hole, using a good word of the form $z = x\diamond x$

1: Build $G = G_3(n)$
2: Select a good word $z = x\diamond x$ with $|x| = n-1$ and $M_z(n) = 3n$
3: Compute $S = \mathrm{Sub}_z(n)$
4: Select some edge e connecting x and some vertex y in G (for $a \in \Sigma$, we have either $e = xa$ where e connects x to y, or $e = ax$ where e connects y to x)
5: Create graph G' from G by deleting the edges in the set $S\backslash\{e\}$ along with any resulting isolated vertices $v \neq x$, and add an edge from vertex x to vertex x labelled by z
6: Find an Eulerian path p in G' (from x to y, if $e = xa$, or from y to x, if $e = ax$)
7: **return** p

Algorithm 3 is used to build a ternary de Bruijn partial word with one hole for a good word of the form $z = x\diamond y$.

Algorithm 3 Constructing a 3-ary de Bruijn word of order $n \geq 2$ with one hole, using a good word of the form $z = x \diamond y$, $x \neq y$

1: Build $G = G_3(n)$
2: Select a good word $z = x \diamond y$ with $|x| = |y| = n - 1$ and $M_z(n) = 3n - 1$
3: Compute $S = \mathrm{Sub}_z(n)$
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 II.8. We construct a 3-ary de Bruijn partial word of order 3 with one hole using Algorithm 3. First, we build the graph $G = G_3(3)$ (see Figure 9). Next, we select good word $z = x \diamond y = 01 \diamond 10$ such that $M_z(3) = 3(3) - 1 = 8$. When we compute $S = \mathrm{Sub}_z(n) = \{010, 011, 012, 101, 111, 121, 110, 210\}$, we find a repeated subword of length 3, 010. Next, we create G' by removing the edges corresponding to subwords in S from the graph G and adding the new edge for z connecting vertices x and y (see Figure 10). Finally, we find an Eulerian path from vertex y to vertex x. One such path gives the de Bruijn partial word: $10002221122001 \diamond 102021201$. Figure 10 shows our path before travelling the edge corresponding to our good word z and our path after the edge for z.

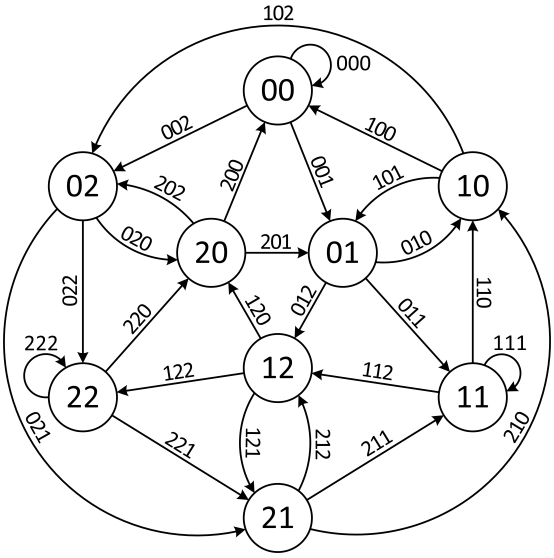

Figure 9. $G_3(3)$

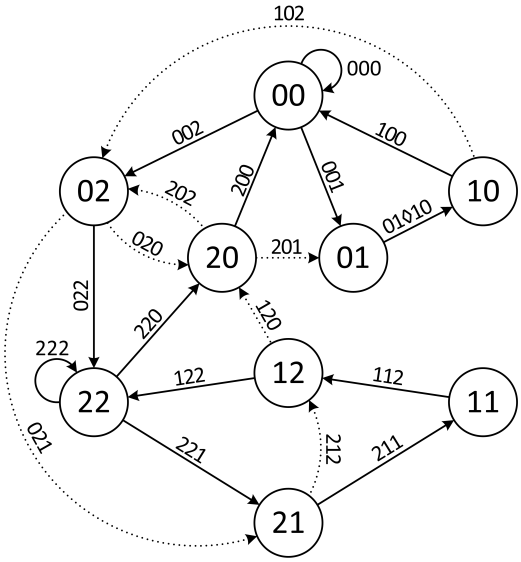

Figure 10. $G_3(3, 01\diamond 10)$

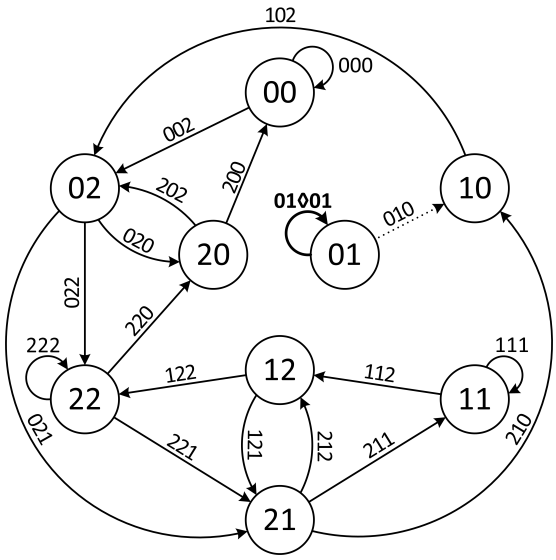

Figure 11. $G_3(3, 01\diamond 01)$ with added edge 010

2.3 The k-ary one-hole case

Using good words of the form $z = x\diamond x$, we can extend these results to $L_k(n, 1)$ for the k-letter alphabet for $n \geq 3$ using Algorithm 4.

Algorithm 4 Constructing a k-ary de Bruijn word of order $n \geq 3$ with one hole, using a good word of the form $z = x\diamond x$

1: Build $G = G_k(n)$
2: Select a good word $z = x\diamond x$ with $|x| = n - 1$ and $M_z(n) = kn$
3: Compute $S = \mathrm{Sub}_z(n)$
4: Select some edge e connecting x and some vertex y in G (for $a \in \Sigma$, we have either $e = xa$ where e connects x to y, or $e = ax$ where e connects y to x)
5: Create graph G' from G by deleting the edges in the set $S\backslash\{e\}$ along with any resulting isolated vertices $v \neq x$, and add an edge from vertex x to vertex x labelled by z
6: Find an Eulerian path p in G' (from x to y, if $e = xa$, or from y to x, if $e = ax$)
7: **return** p

This gives us the following result for the length of k-ary de Bruijn partial words with one hole for $k \geq 3$.

Theorem II.9 ([8]). *For $k \geq 3$ and $n \geq 3$, we have $L_k(n,1) = k^n - (k-2)n$.*

Proof. For any $k \geq 3$ and $n \geq 3$, we can find a good word of the form $z = x \diamond x$ with the maximum number of subwords of length n. Letting $z = x \diamond x = 0^{n-2}1 \diamond 0^{n-2}1$, the first $n-1$ factors of length n begin with a different number of 0's, and are thus distinct. The remaining factor ends with 01, making it distinct from any preceeding factors.

Let G' be the graph constructed using Algorithm 4. To determine if G' has an Eulerian path, first consider node x and its neighbor. If the connecting edge begins at x, then $\texttt{ideg}(x) = 1$ and $\texttt{odeg}(x) = 2$, and $\texttt{ideg}(y) = k$ and $\texttt{odeg}(y) = k-1$. Likewise, if it ends at x, then $\texttt{ideg}(x) = 2$ and $\texttt{odeg}(x) = 1$, and $\texttt{ideg}(y) = k$ and $\texttt{odeg}(y) = k-1$.

Next, consider all vertices other than x and y. Some vertices have the same degree as their counterparts in G, so they are balanced. Others have a lesser degree than their G counterparts. These represent the subwords of length $n-1$ that are compatible with factors of z with one hole, or factors in the multiset $F_1(z, n-1)$. Each occurrence of such factor in z is associated with two factors of length n, once with the factor as a prefix and once with the factor as a suffix. This means that, for any of the vertices in this group, an equal number of incoming and outgoing edges were removed. These all remain balanced. Therefore, G' satisfies Lemma II.1, with an Eulerian path either starting or ending at x.

To derive the length of a de Bruijn partial word represented by G', we start with $k^n + n - 1$, the length of a de Bruijn full word corresponding to G. We subtract the

number of edges removed from G when constructing G', or kn. We add n, which is the number of additional characters needed to add the edge corresponding to $x \diamond x$. Finally, we add one for the additional edge reconnecting x to the remainder of G'. This gives $L_k(n,1) = (k^n + n - 1) - kn + n + 1 = k^n - (k-2)n$. □

2.4 Counting binary de Bruijn partial words with one hole

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 $G_k(n)$. In the binary one-hole case, 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 II.10 ([14]). *Let $G = (V, E)$ be an Eulerian digraph, and let L_G denote the Laplacian matrix of G defined as follows: for $i = j$, $L_G(i,j) = \mathtt{odeg}(v_i) - e$, 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*

$$C \prod_{v \in V} (\mathtt{odeg}(v) - 1)! = C \prod_{v \in V} (\mathtt{ideg}(v) - 1)! \tag{II.1}$$

with C any cofactor of L_G.

To compute $N_2(n,1)$, we need to modify Theorem II.10, since we want to count the number of Eulerian paths.

Theorem II.11 ([8]). *Let $G = (V, E)$ be a digraph, and let $x, y \in V$ be such that* $\texttt{odeg}(x) = \texttt{ideg}(x) + 1$ *and* $\texttt{ideg}(y) = \texttt{odeg}(y) + 1$. *Suppose that G satisfies the conditions of Lemma* II.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 (*II.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 II.10 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 5, to count the number of 2-ary de Bruijn partial words of order n with one hole.

Algorithm 5 ([8]) Computing the number $N_2(n, 1)$, where $n \geq 4$

1: Find the set Z of all good z's of the form $x \diamond y$ 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. 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 II.2, x must be the last vertex of the path because $\texttt{ideg}(x) = \texttt{odeg}(x) + 1$. In step 5, the $(\texttt{ideg}(x) - 1)!$ multiplicative factor is always 1 since $\texttt{ideg}(x) = 2$. 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 II.12. Returning to Example II.3 with $k = 2$ and $n = 4$, up to reversal and renaming of letters, we only need to consider $z_1 = 001\diamond110$ to compute $N_2(n, 1)$. Referring to the graph on the left in Figure 6, $L_{G_2(4,z_1)}$ is shown in Figure 12. Note that the rows and columns corresponding to the vertices 011 and 111 have been removed since all their entries are zeros. If we remove the row and column of vertex 001, we get a determinant of 4. So there are 4 Eulerian paths from 110 to 001 in $G_2(4, z_1)$:

$$11001\diamond110100001, 1100001\diamond1101001, 1101001\diamond1100001, 110100001\diamond11001$$

Since the only candidate words that are good words are $001\diamond110$, its reversal, and their renamings, we get $N_2(4, 1) = 4 \times 4 = 16$.

	000	001	010	100	101	110
000	1	−1	0	0	0	0
001	0	1	0	0	0	−1
010	0	0	1	−1	0	0
100	−1	−1	0	2	0	0
101	0	0	−1	0	1	0
110	0	0	0	−1	−1	2

Figure 12. $L_{G_2(4,z_1)}$

CHAPTER III

DE BRUIJN PARTIAL WORDS WITH TWO OR MORE HOLES

In this chapter, we extend the method of constructing de Bruijn partial words with one hole to include de Bruijn partial words with $h \geq 2$ holes. We then present results for some known lengths of de Bruijn partial words.

3.1 Minimal partial words of maximal subword complexity

As shown in the binary one-hole case, we can find a k-ary de Bruijn partial word of order n with h holes by modifying $G_k(n)$ as follows: we select a good word $z = x\diamond y$, we remove all edges from $G_k(n)$ corresponding to words in $\mathrm{Sub}_z(n)$, and then add an edge z from vertex x to vertex y.

When $h > 1$, if we use this approach, G' is often neither weakly-connected nor weakly-balanced. Thus, we must build an Eulerian path by adding edges. This is seen in the k-ary one-hole case for $k \geq 3$, but as we increase the number of holes, more edges must be added to build an Eulerian path in G'. For example, $G' = G_2(6, 00110\diamond\diamond\diamond10001)$ is neither weakly-balanced nor weakly-connected (Fig. 13). Therefore, we must first reconnect G'. This requires two edges: after adding 110101 and 1100100, G' is weakly-connected but not weakly-balanced. We can weakly-balance G' by adding edge 11011100, after which G' has an Eulerian path corresponding to a de Bruijn partial word. The binary de Bruijn partial word

$$0100000011111101110010011 0\diamond\diamond\diamond 100010111010101 \qquad \text{(III.1)}$$

of length 44 corresponds to the Eulerian path from 01000 to 10101 in G'. A de Bruijn full sequence with the parameters $n = 6, k = 2$ has length 69, much longer than the 44 length sequence that uses three wildcards.

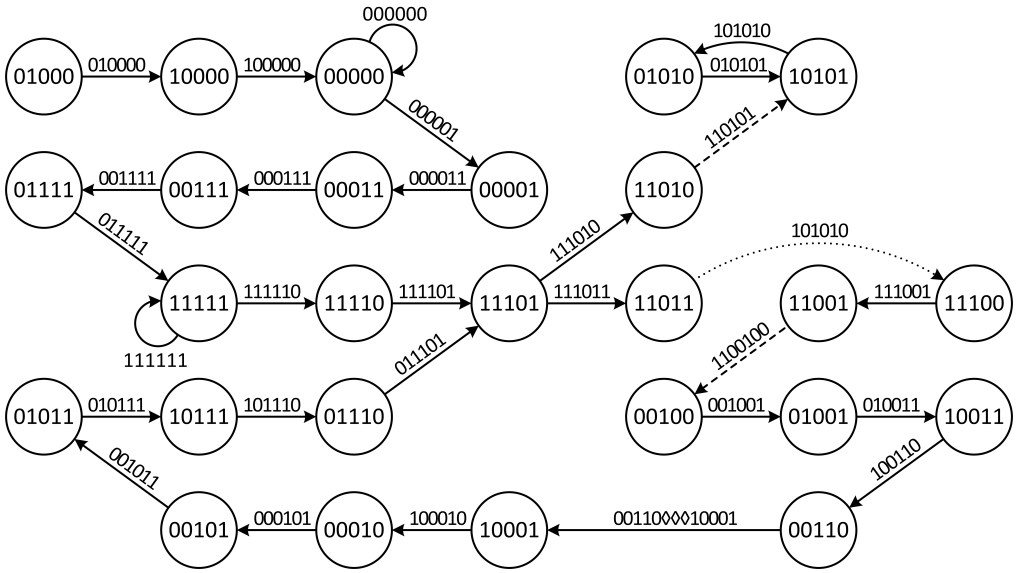

Figure 13. $G_2(6, 00110\diamond\diamond\diamond 10001)$

This allows us to generalize our set of construction algorithms to a single algorithm that applies to all cases for $h \geq 1$, $k \geq 2$, $n \geq 3$. However, we have another problem: for a given h, k, and n, how do we find a good word z? To address this, we can write our algorithm for the general case to apply to any candidate word z, so that the resulting constructed word corresponding to an Eulerian path is a 2^k-complex word of minimal length with h holes in which z appears as a factor. Furthermore, this can help us with the problem of discovering good words by comparing results for the algorithm over the set of all possible candidates. Those producing the results of shortest length are de Bruijn partial words.

Algorithm 6 ([6]) Constructing a minimal k^n-complex partial word with h holes containing a candidate word of the form $z = xz'y$ as a factor

1: Build $G = G_k(n)$
2: Select a candidate word $z = xz'y$ with $|x| = |y| = n - 1$, where the first and last position of z' is a hole and z' contains h holes
3: Compute $S = \mathrm{Sub}_z(n)$
4: Create graph G' from G by deleting the edges in the set S and add an edge from vertex x to vertex y labelled by z
5: Delete any isolated vertices
6: Add the set of edges R, where each edge in R is from a vertex u to a vertex v with label uv', where uv' is the shortest full word with u as a prefix and v as a suffix, such that G' contains an Eulerian path with the smallest possible sum of lengths of edge labels
7: Find an Eulerian path p in G'
8: **return** p

We refer to the resulting graph G' as $G_k(n, z, R)$. We use this algorithm in Section 3.2 to show the construction of some known de Bruijn partial words, and in Section 4.1 we use it to conjecture about the lengths of binary de Bruijn partial words in the $h \geq 2$, $n = h + 3$ case.

3.2 Closed formulas for known lengths of de Bruijn partial words

In this section, we give some closed formulas for the length $L_{k,h}(n)$ of a k-ary de Bruijn partial word with h holes of order n where $h \geq 1$. Firstly, $L_{k,n}(n) = n$ since $\diamond^n$ is a de Bruijn partial word of order n with n holes over a k-letter alphabet, compatible with all words of length n over the k letters. For the binary $(n-1)$- and $(n-2)$-hole cases, we have the following.

Theorem III.1 ([6]). *Let h be a non-negative integer and let $n = h + 1$. Then (1) the partial word $\diamond^h 01^n$ is 2^n-complex and (2) each factor of length $2n$ of $0^n 1 \diamond^h 01^n$ is 2^n-complex. Consequently, $L_{2,n-1}(n) = 2n$ and the only binary de Bruijn partial words of order n with $n-1$ holes are the factors of length $2n$ of $0^n 1 \diamond^{n-1} 01^n$ and $1^n 0 \diamond^{n-1} 10^n$.*

Similarly, for $h \geq 1$, $L_{2,n-2}(n) = 3n-1$. Furthermore, each of those partial words of length $3n-1$ is constructed with a good word that is a factor of either $0^{n-1} 1 \diamond^h 10^{n-1}$ or $1^{n-1} 0 \diamond^h 01^{n-1}$.

Proof. For (1), $\diamond^h 01^n$ has 2^n distinct subwords of length $n = h+1$. Those 2^h that end with 0 are compatible with the factor $\diamond^h 0$, those $1 + \cdots + 2^{h-1}$ ending with 01^i, where $1 \leq i \leq h$, are compatible with $\diamond^{h-i} 01^i$, and 1^n appears as a factor, for a total of 2^n subwords. For (2), set $w = 0^n 1 \diamond^h 01^n$. The prefix of length $2n$ is $0^n 1 \diamond^h$ and the suffix of same length is its reverse complement, and so these factors are 2^n-complex by (1). For $1 \leq i \leq n$, consider the factor $u_i = 0^{n-i} 1 \diamond^{n-1} 01^{i-1}$ that starts at position i of w. It has all words of length n that start with 1 or end with 0 as subwords, since they are compatible with $1 \diamond^{n-1}$ or $\diamond^{n-1} 0$, respectively. Now, consider a word of length n that starts with 0 and ends with 1. Either it has a prefix of the form $0^j 1$ for some $1 \leq j \leq n-i$, compatible with $0^j 1 \diamond^{n-j-1}$, or it has a suffix of the form 01^j for some $1 \leq j \leq i-1$, compatible with $\diamond^{n-j-1} 01^j$. Consequently, $L_{2,n-1}(n) \leq 2n$ holds and it is not difficult to see that $L_{2,n-1}(n) \geq 2n$ also holds. □

Thirdly, we show that $L_{2,2}(n) = 2^n - 2n + 2$ for $n \geq 6$. We first show that $2^n - 2n + 2$ is an upper bound. For $n \geq 6$, we define a digraph G'_n as follows: Start with the de Bruijn graph $G_n = G_2(n)$. Trim G_n by deleting all edges that are in

$\mathrm{Sub}_z(n)$, where $z = 0110^{n-4}\diamond\diamond 10^{n-2}$. Then, add a new edge from 0110^{n-4} to 10^{n-2} labelled by z. Remove any isolated vertices.

Lemma III.2 ([6]). *For $n \geq 6$, if G'_n is weakly-connected, then G'_n has an Eulerian path.*

Proof. Using Lemma II.1, it is sufficient to show that there exist vertices x, y in G'_n such that $\texttt{odeg}(x) = \texttt{ideg}(x) + 1$ and $\texttt{ideg}(y) = \texttt{odeg}(y) + 1$, and for all vertices $v \notin \{x, y\}$, $\texttt{ideg}(v) = \texttt{odeg}(v)$. Note that $z = 0110^{n-4}\diamond\diamond 10^{n-2}$ satisfies $p_z(n) = 4n - 2$, with repeated subwords 0110^{n-3} and 110^{n-2}. Then, observe that 0110^{n-4} and 10^{n-2} have degree $(1, 1)$. Now suppose that v is a subword of z of length $n - 1$ not equal to $0110^{n-4}, 10^{n-2}, 110^{n-4}1$, or 010^{n-3}. Note that v is a prefix of some v_1 and a suffix of some v_2, where v_1 and v_2 are subwords of length n of z. Therefore, v remains balanced (e.g., $\texttt{ideg}(v) = \texttt{odeg}(v)$). Finally, vertices $110^{n-4}1$ and 010^{n-3} have degree $(1, 0)$ and $(0, 1)$, respectively. □

Lemma III.3 ([6]). *For $n \geq 6$, G'_n is weakly-connected.*

Proof. We show that there exists a path from any vertex u in G'_n to $v = 110^{n-4}1$. First, $11110^{n-4}, 1110^{n-4}1 \notin \mathrm{Sub}_z(n)$, so it is sufficient to show that there exists a path from u to $v' = 11110^{n-5}$. Since z avoids 1^4, if u ends in 111 we are done, as $u[0..n-4)1^4, u[1..n-4)1^40, \ldots, 1^40^{n-4}$ is a valid path to v'. So, if we can reach a vertex ending in 111 we are done. Now, suppose that u ends in 11, and that $u1 \in \mathrm{Sub}_z(n)$ (so there is no edge from u to a vertex ending in 111). This implies $u = 10^{n-4}11$. However, $10^{n-4}110, 0^{n-4}1101, 0^{n-5}11011, 0^{n-6}110111$ provides a valid path to a vertex ending in 111, implying that there exists a path from u to v as desired. So, if we can reach a vertex ending in 11 we are done.

Next, suppose that u ends in 01, and that $u1 \in \mathrm{Sub}_z(n)$ (so there is no edge from u to a vertex ending in 11). This implies $u = v$ or $u = 10^{n-3}1$. In this case, $10^{n-3}10, 0^{n-3}101, 0^{n-4}1011$ provides a valid path to a vertex ending in 1. Therefore, if we can reach a vertex ending in 1 we are done. Furthermore, $0^{n-1}1 \notin \mathrm{Sub}_z(n)$, so there exists a path from 0^{n-1} to v. Last, suppose that u ends in 0. If $u1 \notin \mathrm{Sub}_z(n)$ we are done. Furthermore, if $u0, u1 \in \mathrm{Sub}_z(n)$, we have that $\mathtt{odeg}(u) = 0$, so that $u = v$ or u is isolated. Thus, if $u \neq v$ and u is not isolated, u is connected to a vertex ending in 00. Repeating the above argument, we get that either u is connected to v, or u is connected to $u^{n-2} = 0^{n-1}$. In either case, u is connected to v. □

Lemma III.4 ([6]). *Let $u = x\diamond w\diamond y$, where x, w, y are full words, be 2^n-complex. Then (1) $|u| \geq 2^n - 2n + |w|$, with equality only if u has no repeated subwords and (2) if $|x| = n - 2 - |w| - p$ (similarly, $|y| = n - 2 - |w| - p$) for $p \geq 0$, then $|u| \geq 2^n - 2n + 3p + |w|$, with equality only if u has no repeated subwords.*

Proof. Note that $|F_0(u,n)| = |u| - n + 1 - |F_1(u,n)| - |F_2(u,n)|$. For (1), looking at the positions of the holes in u, we have $|F_1(u,n)| \leq 2|w| + 2$ and $|F_2(u,n)| \leq n - |w| - 1$. Notice that if $|w| \geq n$, then $|F_2(u,n)| = 0$. Thus, $2^n \leq |u| - n + 1 + |F_1(u,n)| + 3|F_2(u,n)| \leq |u| + 2n - |w|$. For (2), we have $|F_1(u,n)| \leq 2|w| + 1$ and $|F_2(u,n)| \leq n - 1 - p - |w|$. Thus, $2^n \leq |u| - n + 1 + |F_1(u,n)| + 3|F_2(u,n)| \leq |u| + 2n - 3p - |w| - 1$. □

Theorem III.5 ([6]). *For $n \geq 6$, $L_{2,2}(n) = 2^n - 2n + 2$.*

Proof. From Lemmas III.2 and III.3, there exists an Eulerian path in G'_n. Traversing it generates a word w of length $2^n - p_z(n) + |z| = 2^n - 4n + 2 + 2n = 2^n - 2n + 2$ with $p_w(n) = 2^n$, and so $L_{2,2}(n) \leq 2^n - 2n + 2$. Now let $u = x\diamond w\diamond y$ be 2^n-complex, where x, w, y are full. By Lemma III.4, if $|u| < 2^n - 2n + 2$, we have $|w| \leq 1$ and

$|x|, |y| \geq n-2-|w|$. Furthermore, [8, Corollary 2] tells us $|u| \geq 2^n - 2n$, so for $n \geq 6$, if $|x| = n-2-|w|$, then $|y| \geq n-1-|w|$, and similarly, if $|y| = n-2-|w|$, then $|x| \geq n-1-|w|$. Let a, b be the letters of the alphabet.

First, assume that $|x| = n-2-|w|$ (the case where $|y| = n-2-|w|$ is similar). Thus, $|y| \geq n-1-|w|$ and by Lemma III.4(2), $|u| < 2^n - 2n + 2$ only if $|w| = 0$ and u has no repeated subwords. Therefore, we can write $u = x \diamond\diamond \tilde{y}cy'$, where $y = \tilde{y}cy'$, $c \in \{a, b\}$, y' is some (perhaps empty) full word and $|\tilde{y}| = n-2$. Now let $\tilde{y}_a = a\tilde{y}$ and $\tilde{y}_b = b\tilde{y}$. Note that $\tilde{y}_a c, \tilde{y}_b c \in \mathrm{Sub}_u(n)$. Furthermore, $\tilde{y}_a \overline{c}, \tilde{y}_b \overline{c}$ are also full words of length n, where $\overline{c}$ is complement of letter c, and thus must appear starting at a different index in u as well. At most one of $\tilde{y}_a \overline{c}, \tilde{y}_b \overline{c}$ is compatible with a prefix of u. Assume without loss of generality that $\tilde{y}_a \overline{c}$ is not compatible with a prefix of u. Thus, $e\tilde{y}_a \overline{c}$, for some letter e, appears somewhere implying that $e\tilde{y}_a \in \mathrm{Sub}_u(n)$ is a repeated subword, a contradiction.

Next, assume that both $|x|, |y| \geq n-1-|w|$. By Lemma III.4(1), $|w| < 2$ and $|u| < 2^n - 2n + 2$ only if u has zero or one repeated subwords. We show that u contains at least two repeated subwords. We can write $u = x' d\tilde{x} \diamond w \diamond \tilde{y}cy'$, where $x = x' d\tilde{x}$, $y = \tilde{y}cy'$, $c, d \in \{a, b\}$, x', y' are some (perhaps empty) full words and $|\tilde{x}| = |\tilde{y}| = n-2-|w|$. Similar to above, let $\tilde{x}_a = \tilde{x}aw, \tilde{x}_b = \tilde{x}bw, \tilde{y}_a = wa\tilde{y}$, and $\tilde{y}_b = wb\tilde{y}$. All of $d\tilde{x}_a, d\tilde{x}_b, \tilde{y}_a c, \tilde{y}_b c \in \mathrm{Sub}_u(n)$. By similar reasoning to above, without loss of generality $\overline{d}\tilde{x}_a$ appears (not compatible with a suffix) somewhere else in u. This implies that $\tilde{x}_a f$ is a repeated subword for some letter f. Likewise, $e\tilde{y}_a$ is also a repeated subword for some letter e, implying that $|u| \geq 2^n - 2n + 2$. This case is summarized below:

$$u = \quad x'_0 \cdot \cdots \cdot x'_{i-1} d \overbrace{\tilde{x}_0 \cdot \cdots \cdot \tilde{x}_{n-3-|w|} \diamond w}^{\tilde{x}_a, \tilde{x}_b} \diamond \tilde{y}_0 \cdot \cdots \cdot \overbrace{\tilde{y}_{n-3-|w|} c y'_0 \cdot \cdots}^{\overline{d}\tilde{x}_a f} \cdot y'_{j-1}$$

$$u = \quad x'_0 \cdot \overbrace{\cdots \cdot x'_{i-1} d \tilde{x}_0 \cdot \cdots}^{e\tilde{y}_a \overline{c}} \cdot \tilde{x}_{n-3-|w|} \diamond \overbrace{w \diamond \tilde{y}_0 \cdot \cdots \cdot \tilde{y}_{n-3-|w|}}^{\tilde{y}_a, \tilde{y}_b} c y'_0 \cdot \cdots \cdot y'_{j-1}$$

Thus, $|u| \geq 2^n - 2n + 2$, so $L_{2,2}(n) \geq 2^n - 2n + 2$. □

CHAPTER IV

DISCUSSION

In this chapter, we offer conjecture for the length of binary de Bruijn partial words in the case where $h \geq 2$, $n = h + 3$ and present open questions.

4.1 Conjecture for the binary $h \geq 3, n = h + 3$ case

There is evidence suggesting that in the $h \geq 2$, $n = h+3$ case, there exist 2-ary de Bruijn partial words that can be constructed from good words z that follow a pattern.

Conjecture IV.1. *Let $h \geq 2$ and let $n = h+3$. Then a 2-ary de Bruijn partial word w of order n with h holes can be built by first selecting a good word z as follows:*

$$z = \begin{cases} 0001\diamond\diamond0110 & \text{if } n = 5; \\ 0^{n-2}1\diamond^h01^{\frac{n-2}{2}}0^{\frac{n-2}{2}} & \text{if } n \text{ is even;} \\ 0^{n-2}1\diamond^h01^{\frac{n-3}{2}}0^{\frac{n-1}{2}} & \text{otherwise.} \end{cases}$$

Then w corresponds to an Eulerian path in $G = G_2(n, z, R)$ constructed as described in Algorithm 6.

Consider $z = 0001\diamond\diamond0110$ for $h = 2$, $n = 5$. Let $G' = G_2(5, 0001\diamond\diamond0110)$. Then we find a single Eulerian path from 1011 to 0010 in G', which corresponds to the path label 101111000001$\diamond\diamond$011010010, a binary de Bruijn partial word of order 5 with 2 holes (see Figure 14).

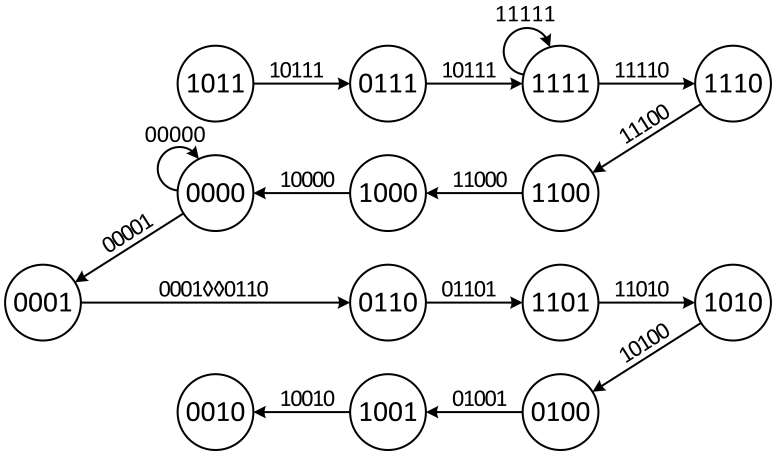

Figure 14. $G_2(5, 0001\diamond\diamond 0110)$

For $h = 3$, $n = 6$, we take $z = 00001\diamond\diamond\diamond 01100$. The resulting graph $G_2(6, z)$ does not have an Eulerian path; we build an Eulerian path with a label of minimal length by adding edges 0101011, 0001010, and 0010011 (see Figure 15) and call the resulting graph G'. We observe G' contains the path

$$10111 \overset{101111}{\longrightarrow} 01111 \overset{011111}{\longrightarrow} 11111 \overset{111111}{\longrightarrow} 11111 \overset{111110}{\longrightarrow} 11110 \overset{111100}{\longrightarrow} 11100 \overset{111000}{\longrightarrow} 11000$$
$$\overset{110000}{\longrightarrow} 10000 \overset{100000}{\longrightarrow} 00000 \overset{000000}{\longrightarrow} 00000 \overset{000001}{\longrightarrow} 00001 \overset{00001\diamond\diamond\diamond 01100}{\longrightarrow} 01100 \overset{011001}{\longrightarrow} 11001$$
$$\overset{110010}{\longrightarrow} 10010 \overset{100100}{\longrightarrow} 00100$$

which is similar to the Eulerian path in the graph in Figure 14. Both paths start in node 101^{n-3}; both traverse 01^{n-1}, then 1^n, followed by a sequence of edges $1^{n-i}0^i$ for $1 \leq i \leq n$, then $0^{n-1}1$ and z.

Note that most vertices in G' have both an indegree and an outdegree of 1. When such a vertex is reached in a path in G', there is only a single possibility for the next traversed edge. Furthermore, when 0^{n-1} (or 1^{n-1}) is first reached in an Eulerian

path, 0^n (or 1^n) must be the very next edge travelled, followed by $0^{n-1}1$ (or $1^{n-1}0$). This allows us to reduce the size of our digraph by replacing each longest sequence of edges forming an "unavoidable" path in G' from vertex u to vertex v with a single edge (u, v) and removing isolated vertices. For example, the path above corresponds to label 101111110000001$\diamond\diamond\diamond$01100100; we can add an edge from 10111 to 00100 with this label and remove all of the edges belonging to the path from our graph, along with any vertices that become isolated (see Figure 16). Thus, we have a condensed graph equivalent to G'. Figures 17 and 18 for the $h = 4, n = 7$ and $h = 5, n = 8$ cases are reduced in this way for clarity (otherwise, they contain too many vertices to draw clearly). All edges added to weakly-connect or weakly-balance a digraph are shown. Also note that each separate component of G' existing before edges 0101011, 0001010, and 0010011 are added is enclosed in a grey box.

For $h = 4$, $n = 7$, we have:

$$0110111001000111001110111101000010110101001010111111$$
$$11100000001\diamond\diamond\diamond\diamond0110001001111$$

of length 78 for the potentially good word $z = 000001\diamond\diamond\diamond\diamond011000$ (compared to length 134 for a binary de Bruijn full word of order 7), and for $h = 5$, $n = 8$, we have:

$$10101111111100000000 1\diamond\diamond\diamond\diamond\diamond0111000100011110011000 11$$
$$00111101100001001111101111001000011011111010000010$$
$$11101010001010101001101011001010010110110100 10010$$

of length 149 for the potentially good word $z = 0000001\diamond\diamond\diamond\diamond\diamond 0111000$ (compared to length 263 for a binary de Bruijn full word of order 8). While no shorter 2^7 or 2^8-complex words have been discovered in the course of this research, these have not yet been verified as de Bruijn partial words.

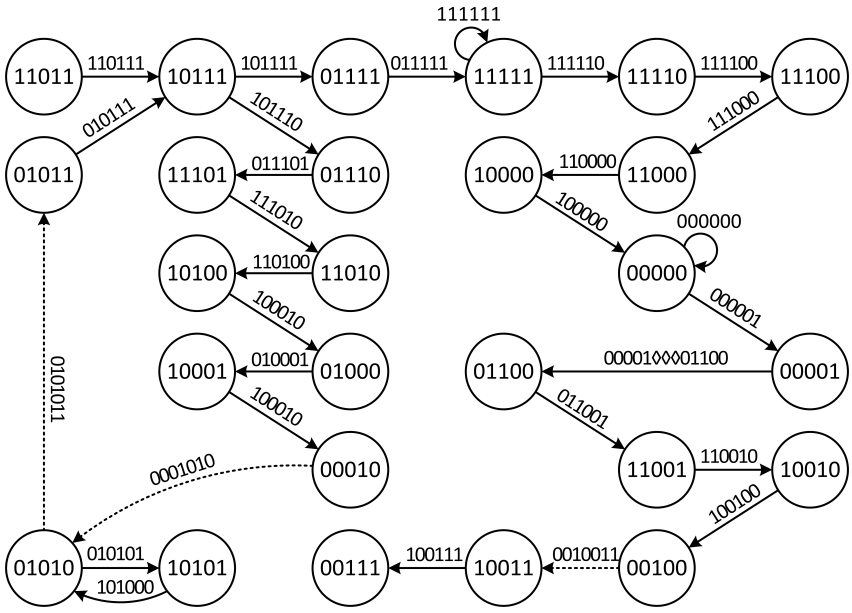

Figure 15. $G_2(6, 00001\diamond\diamond\diamond 01100)$ with added edges

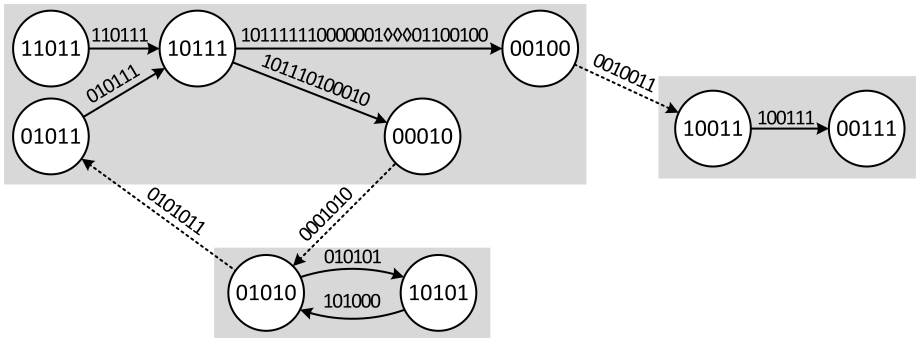

Figure 16. $G_2(6, 00001\diamond\diamond\diamond 01100)$ with added edges, condensed

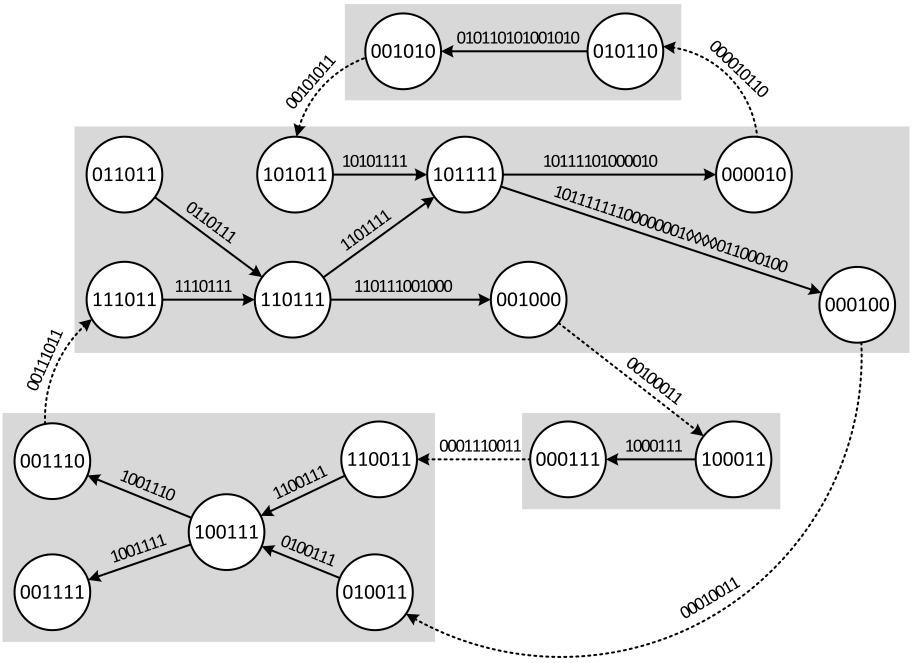

Figure 17. $G_2(7, 000001\diamond\diamond\diamond\diamond011000)$ with added edges, condensed

4.2 Conclusion

The question of binary de Bruijn partial word length has been determined for the one- and two-hole cases, along with cases for $n-1$- and $n-2$-hole cases. A de Bruijn partial word can be constructed for each of these cases in linear time. The question of length is still open for binary de Bruijn partial words in the general case for $h \geq 2$. Figure 19 summarizes the findings of known 2^n-complex words of minimal lengths for cases of n and h, with figures in bold representing lengths for confirmed de Bruijn partial words. Binary de Bruijn partial words with one-hole can be counted using Laplacian matrices, and in the $n-1$- and $n-2$-hole cases, they can be counted by counting factors of the special words constructed in Theorem III.1.

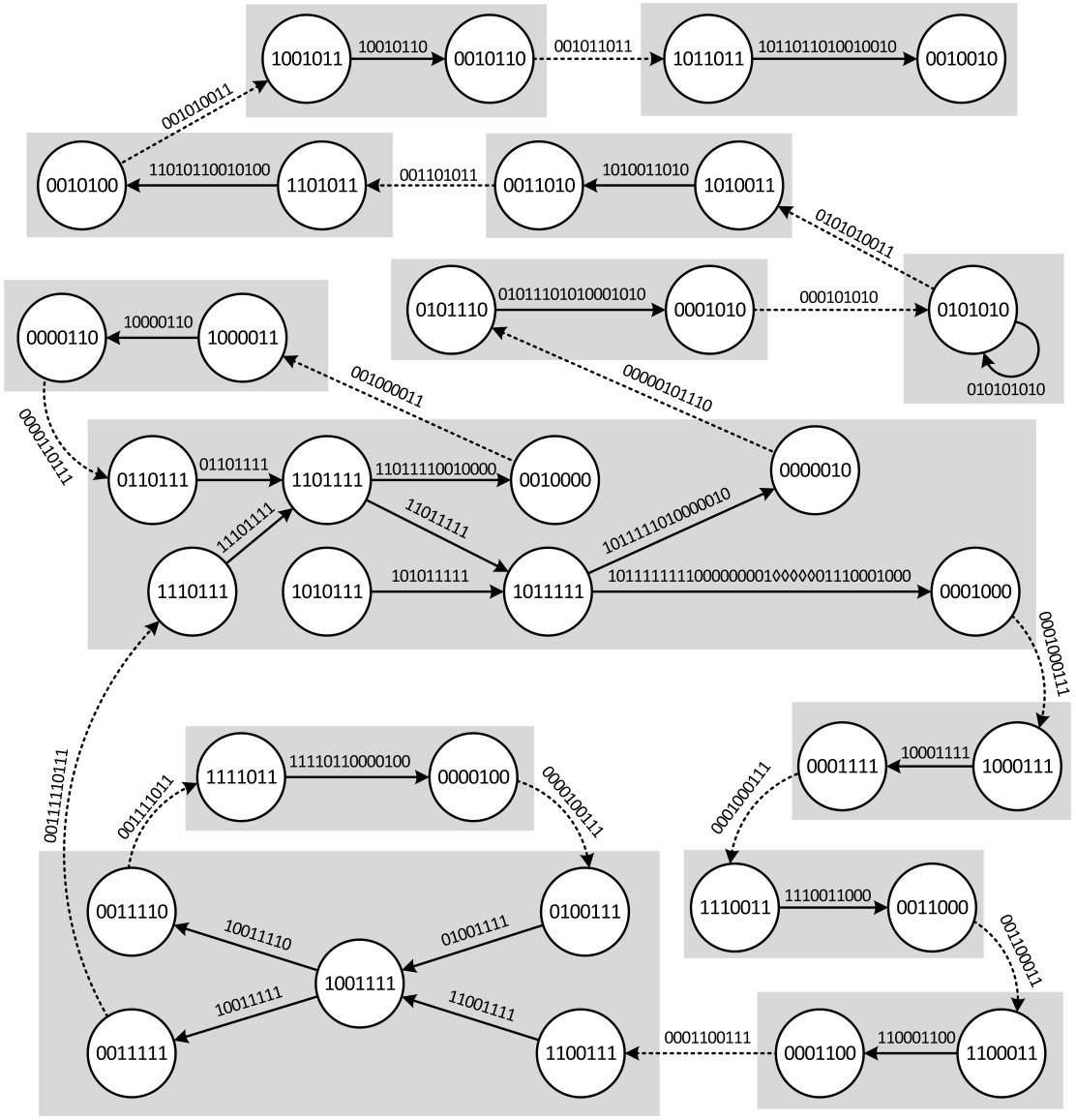

Figure 18. $G_2(8, 0000001\diamond\diamond\diamond\diamond\diamond0111000)$ with added edges, condensed

For the k-ary case with $k \geq 3$, the length of a k-ary de Bruijn partial word of order n with h holes is known for the one-hole case and can be constructed efficiently. The length of k-ary de Bruijn partial words with $h \geq 2$ is an open question. Counting de Bruijn partial words is an open question for all $k \geq 3$.

Order \ Number of Holes	0	1	2	3	4	5	6	7	8
0	0	1	2	3	4	5	6	7	8
1	2	1	2	3	4	5	6	7	8
2	5	4	2	3	4	5	6	7	8
3	10	8	6	3	4	5	6	7	8
4	19	15	11	8	4	5	6	7	8
5	36	31	24	14	10	5	6	7	8
6	69	63	54	44	17	12	6	7	8
7	134	127	116	102	78	20	14	7	8
8	263	255	242	225	210	149	23	16	8

Figure 19. Some $L_{2,h(n)}$ lengths

REFERENCES

[1] M. A. Alekseyev and P. A. Pevzner. Colored de Bruijn graphs and the genome halving problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(1):98–107, 2007.

[2] J.-P. Allouche. Sur la complexité des suites infinies. *Bulletin of the Belgian Mathematical Society*, 1:133–143, 1994.

[3] J.-P. Allouche and J. Shallit. *Automatic Sequences: Theory, Applications, Generalizations*. Cambridge University Press, 2003.

[4] B. Blakeley, F. Blanchet-Sadri, J. Gunter, and N. Rampersad. On the complexity of deciding avoidability of sets of partial words. *Theoretical Computer Science*, 411:4263–4271, 2010. (`www.uncg.edu/cmp/research/unavoidablesets3`).

[5] F. Blanchet-Sadri. *Algorithmic Combinatorics on Partial Words*. Chapman & Hall/CRC Press, Boca Raton, FL, 2008.

[6] F. Blanchet-Sadri, D. Allums, J. Lensmire, and B. J. Wyatt. Constructing minimal partial words of maximum subword complexity. In *JM 2012, 14th Mons Days of Theoretical Computer Science, September 11-14, 2012, Universite catholique de Louvain, Belgium*, Universite catholique de Louvain, Belgium. (`www.uncg.edu/cmp/research/subwordcomplexity3`).

[7] F. Blanchet-Sadri and J. Lensmire. On minimal Sturmian partial words. *Discrete Applied Mathematics*, 159(8):733–745, 2011.

[8] F. Blanchet-Sadri, J. Schwartz, S. Stich, and B.J. Wyatt. Binary de Bruijn partial words with one hole. In J. Kratochvil et al., editor, *TAMC 2010, 7th Annual Conference on Theory and Applications of Models of Computation, Prague, Czech Republic*, volume 6108 of *Lecture Notes in Computer Science*, pages 128–138, Berlin, Heidelberg, 2010. Springer-Verlag. (`www.uncg.edu/cmp/research/subwordcomplexity`).

[9] J. Cassaigne. Complexité et facteurs spéciaux. *Bulletin of the Belgian Mathematical Society*, 4(1):67–88, 1997.

[10] N. G. De Bruijn. 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.

[11] S. Ferenczi. Complexity of sequences and dynamical systems. *Discrete Mathematics*, 206:145–154, 1999.

[12] I. Gheorghiciuc. The subword complexity of a class of infinite binary words. *Advances in Applied Mathematics*, 39:237–259, 2007.

[13] J. L. Gross and J. Yellen. *Handbook of Graph Theory*. CRC Press, 2004.

[14] R. P. Stanley. *Enumerative Combinatorics*, volume 2. Cambridge University Press, Cambridge, 2001.