Generalized De Bruijn Words, Invertible Necklaces, and the Burrows–Wheeler Transform

Gabriele Fici
Dipartimento di Matematica e Informatica, Università di Palermo, Italy

Estéban Gabory
Dipartimento di Matematica e Informatica, Università di Palermo, Italy

Abstract

We define generalized de Bruijn words as those words having a Burrows–Wheeler transform that is a concatenation of permutations of the alphabet. We show that generalized de Bruijn words are in 1-to-1 correspondence with Hamiltonian cycles in the generalized de Bruijn graphs, introduced in the early '80s in the context of network design. When the size of the alphabet is a prime p, we define invertible necklaces as those whose BWT-matrix is non-singular. We show that invertible necklaces of length n correspond to normal bases of the finite field $\mathbb{F}_{p^n}$, and that they form an Abelian group isomorphic to the Reutenauer group RG_p^n. Using known results in abstract algebra, we can make a bridge between generalized de Bruijn words and invertible necklaces. In particular, we highlight a correspondence between binary de Bruijn words of order $d+1$, binary necklaces of length 2^d having an odd number of 1's, invertible BWT matrices of size $2^d \times 2^d$, and normal bases of the finite field $\mathbb{F}_{2^{2^d}}$.

2012 ACM Subject Classification Mathematics of computing → Combinatorics on words

Keywords and phrases Burrows–Wheeler Transform, Generalized de Bruijn Word, Generalized de Bruijn Graph, Circulant Matrix, Invertible Necklace, Sandpile Group, Reutenauer Group

Digital Object Identifier 10.4230/LIPIcs.MFCS.2025.48

Funding *Gabriele Fici*: Supported by MIUR project PRIN 2022 APML – 20229BCXNW.
Estéban Gabory: Supported by MIUR project PRIN 2022 APML – 20229BCXNW.

1 Introduction

The Burrows–Wheeler matrix of a word w is the matrix whose rows are the conjugates (rotations) of w in ascending lexicographic order. Its last column is called the Burrows–Wheeler transform of w, and has the property of being easier to compress when the input word is a repetitive text. For this reason, it is largely used in textual data compression, in particular in bioinformatics [21, 19, 18]. But the Burrows–Wheeler transform also has many interesting combinatorial properties (see [31, 12] for a survey). Since the Burrows–Wheeler transform is the same for any conjugate of w, it can be viewed as a map from the set of necklaces (conjugacy classes of words) to the set of words.

A *de Bruijn word* of order d over an alphabet Σ is a necklace of length $|\Sigma|^d$ such that each of the $|\Sigma|^d$ distinct words of length d occurs exactly once in it, as a cyclic factor. Since each of the $|\Sigma|^{d-1}$ distinct words of length $d-1$ occurs as a prefix of $|\Sigma|$ consecutive rows of the Burrows–Wheeler matrix of a de Bruijn word, the Burrows–Wheeler transform of a de Bruijn word is characterized (among words that are images under the Burrows–Wheeler transform) by the fact that it is a concatenation of $|\Sigma|^{d-1}$ alphabet-permutations (de Bruijn words of order 1). This motivates us to define a *generalized de Bruijn word* as one whose Burrows–Wheeler transform is a concatenation of (any number of) alphabet-permutations.

As is well known, de Bruijn words are in 1-to-1 correspondence with Hamiltonian cycles in de Bruijn graphs. We show that, analogously, generalized de Bruijn words are in 1-to-1 correspondence with Hamiltonian cycles in *generalized de Bruijn graphs*, introduced in the


© Gabriele Fici and Estéban Gabory;
licensed under Creative Commons License CC-BY 4.0

early 80's independently by Imase and Itoh [16], and by Reddy, Pradhan, and Kuhl [29] (see also [10]) in the context of network design. The generalized de Bruijn graph $\mathrm{DB}(k, n)$ has vertices $\{0, 1, \ldots, n-1\}$ and for every vertex m there is an edge from m to $km + i \mod (n)$ for every $i = 0, 1, \ldots, k-1$.

In particular, we provide a new and simple interpretation of the well-known correspondence between de Bruijn words and Hamiltonian cycles in de Bruijn graphs in terms of the *inverse standard permutation* of the Burrows–Wheeler transform, and we show that this interpretation extends to the generalized case.

The other class of words we introduce is that of *invertible necklaces*. We call an aperiodic necklace over an alphabet of prime size p invertible if its Burrows–Wheeler matrix is nonsingular, i.e., has determinant nonzero modulo p. We show that each invertible necklace of length n, as a set of vectors over the base field $\mathbb{F}_p$, corresponds to a normal base of the field $\mathbb{F}_{p^n}$. Invertible necklaces can also be represented by conjugacy classes of their circulant matrices, and they form an Abelian group, called the *Reutenauer group* RG_p^n [11]. The set of invertible necklaces of a given length n can therefore be endowed with a multiplication operation that makes it isomorphic to the n-th Reutenauer group.

Using a known result in abstract algebra [7], we show that every aperiodic necklace of length n with non-zero weight modulo p (the weight of a word is the sum of its digits) is invertible if and only if n is either a power of p or a p-rooted prime. However, it is an open problem to decide whether there are infinitely many lengths n, different from a power of p, for which every aperiodic necklace of length n with non-zero weight modulo p has an invertible Burrows–Wheeler matrix. This seems to be a challenging problem, as it is related to Artin's conjecture on primitive roots.

In the last part of the paper, we show a connection between generalized de Bruijn words and invertible necklaces. Chan, Hollmann and Pasechnik [5] proved that for every prime p and any n, one has $RG_p^n \cong K(\mathrm{DB}(p, n)) \oplus \mathbb{Z}_{p-1}$, where $K(G)$ denotes the sandpile group of the graph G. Since the structure of the sandpile group of an Eulerian graph is determined by the invariant factors of its Laplacian matrix, Chan et al. gave the precise structure of sandpile groups of generalized de Bruijn graphs, and hence of Reutenauer groups.

In particular, when $p = 2$ and $n = 2^d$, we show that there is a bijection between de Bruijn words of order d and binary necklaces of length 2^{d-1} with an odd number of 1's.

Our contributions

We introduce two new classes of circular words: generalized de Bruijn words and invertible necklaces, both of which are defined in terms of the Burrows–Wheeler transform. We show that generalized de Bruijn words correspond to those whose Burrows–Wheeler transform has an inverse standard permutation that labels a Hamiltonian cycle in a generalized de Bruijn graph (Theorem 9). This result allows us to derive a formula for counting the number of generalized de Bruijn words using the BEST theorem (Theorem 12).

Next, we focus on alphabets of prime size p. We define invertible necklaces as those whose Burrows–Wheeler matrix is non-singular. We demonstrate that the conjugacy classes of invertible necklaces form an Abelian group, isomorphic to the Reutenauer group of invertible circulant matrices. This leads to a characterization of the equivalence of various properties of invertible necklaces (Theorem 16).

In the final section, we connect generalized de Bruijn words and invertible necklaces by using a previously known isomorphism between the sandpile groups of generalized de Bruijn graphs and Reutenauer groups. We further speculate on the implications of this isomorphism, and end by showing a bijection between binary de Bruijn words, invertible necklaces, and normal bases of finite fields (Theorem 26).

Related work

Several generalizations of de Bruijn words (sequences) have been proposed in the literature [4, 1, 28], all based on the number of occurrences of factors (substrings) of some kind.

The relation between the BWT and de Bruijn words has been studied in several papers. In [14], the author extended the BWT to multisets of necklaces and characterized de Bruijn sets – generalizations of de Bruijn words – by their BWT image, leading to a BWT-based characterization of de Bruijn words when the multiset has a single element. In [24], this correspondence is used to design an efficient algorithm for constructing arbitrary de Bruijn words.

Recently, the algebraic structures generated by invertible BWT matrices have also been explored. In [35, 30], the authors showed that BWT matrices of Christoffel words are closed under multiplication and described the multiplicative group of invertible BWT matrices of Christoffel words. Christoffel words are the unbordered factors of aperiodic infinite words of minimal complexity (Sturmian words) [3]. The conjugacy classes of Christoffel words are precisely the binary aperiodic necklaces whose BWT has the minimal possible number of equal-letter runs [27].

Paper organization

Our paper is structured as follows. In Section 2, we introduce the necessary preliminaries on words, necklaces, and the Burrows–Wheeler transform. Section 3 defines generalized de Bruijn words and establishes their graph-theoretic characterization using generalized de Bruijn graphs, which extends the classical correspondence between de Bruijn words and de Bruijn graphs. In Section 4, we introduce the concept of an invertible necklace and relate it to circulant matrices and Reutenauer groups. Section 5 connects these concepts, using known group isomorphisms to show a bijection between generalized de Bruijn words and invertible necklaces in the context of prime-sized alphabets. Finally, in Section 6, we summarize our findings and propose open problems and future directions.

2 Preliminaries

We begin by introducing some preliminary definitions related to strings and words. For a thorough introduction, we refer the reader to [9], and [25].

Let $\Sigma_k = \{0, 1, \dots, k-1\}$, $k > 1$, be a sorted set of *letters*.

A *word* over the alphabet Σ_k is a concatenation of elements of Σ_k. The *length* of a word w is denoted by $|w|$. For a letter $i \in \Sigma_k$, $|w|_i$ denotes the number of occurrences of i in w. The vector $(|w|_0, \dots, |w|_{k-1})$ is the *Parikh vector* of w.

Let $w = w_0 \cdots w_{n-2} w_{n-1}$ be a word of length $n > 0$. The *weight* of w is $\sum_{j=0}^{n-1} w_j$. When $n > 1$, the *shift* of w is the word $\sigma(w) = w_{n-1} w_0 \cdots w_{n-2}$.

For a word w, the *n-th power* of w is the word w^n obtained by concatenating n copies of w.

▶ **Definition 1.** *We call an* alphabet-permutation *a word over Σ_k that contains each letter of Σ_k exactly once, and an* alphabet-permutation power *a concatenation of one or more alphabet-permutations over Σ_k.*

▶ **Example 2.** The word $w = 210201102102120$ is an alphabet-permutation power over Σ_3.

Notice that an alphabet-permutation power of length n has a balanced Parikh vector, i.e., a Parikh vector of the form $(\frac{n}{k}, \frac{n}{k}, \dots, \frac{n}{k})$.

In the binary case, a word w is an alphabet-permutation power if and only if $w = \tau(v)$, where v is a binary word of length $|w|/2$ and τ is the *Thue–Morse morphism*, the substitution that maps 0 to 01 and 1 to 10.

Two words w and w' are *conjugates* if $w = uv$ and $w' = vu$ for some words u and v. The conjugacy class of a word w can be obtained by repeatedly applying the shift operator, and contains $|w|$ distinct elements if and only if w is *primitive*, i.e., for any nonempty word v and integer n, $w = v^n$ implies $n = 1$. A *necklace* (resp. *aperiodic necklace*) $[w]$ is a conjugacy class of words (resp. of *primitive words*). Necklaces are also called *circular words*. Notice that, given a word w, the weight of w is invariant under shift, hence we can define the *weight* $wt([w])$ of the necklace $[w]$ as the weight of any of its representatives.

For example, $[1100] = \{1100, 0110, 0011, 1001\}$ is an aperiodic necklace, while $[1010] = \{1010, 0101\}$ is not aperiodic.

The number of aperiodic necklaces of length n over Σ_k is given by

$$\overline{\mathrm{N}}(k,n) = \frac{1}{n}\sum_{d|n} \mu\left(\frac{n}{d}\right) k^d$$

where $\mu(n)$ is the Möbius function, defined by: $\mu(1) = 1$, $\mu(n) = (-1)^j$ if n is the product of j distinct primes or 0 otherwise, i.e., if n is divisible by the square of a prime number.

The number of necklaces of length n over Σ_k is given by

$$\mathrm{N}(k,n) = \sum_{d|n} \overline{\mathrm{N}}(k,d) = \frac{1}{n}\sum_{d|n} \phi\left(\frac{n}{d}\right) k^d$$

where $\phi(n)$ is the Euler's totient function. Recall that the Euler's totient function $\phi(n)$ counts the number of positive integers less than or equal to n and coprime with n, i.e., the order of the multiplicative group $\mathbb{Z}_n^*$, and can be computed using the formula

$$\phi(n) = n\prod_{p|n}\left(1 - \frac{1}{p}\right)$$

for $n > 1$, and $\phi(1) = 1$, where the product is taken over the distinct primes dividing n.

The *Burrows–Wheeler matrix* (BWT matrix) of a necklace $[w]$ is the matrix whose rows are the $|w|$ shifts of w in ascending[1] lexicographic order. Let us denote by F and L, respectively, the first and the last column of the BWT matrix of $[w]$. We have that $F = 0^{n_0}1^{n_1}\cdots(k-1)^{n_{k-1}}$, where $(n_0, \ldots, n_{k-1})$ is the Parikh vector of $[w]$; L, instead, is the *Burrows–Wheeler Transform* (BWT) of $[w]$. The BWT is therefore a map from the set of necklaces to the set of words. As is well known, it is an injective map (we describe below how to invert it). A word is a *BWT image* if it is the BWT of some necklace.

The *standard permutation* of a word $u = u_0u_1\cdots u_{n-1}$, $u_i \in \Sigma_k$, is the permutation π_u of $\{0, 1, \ldots, n-1\}$ such that $\pi_u(i) < \pi_u(j)$ if and only if $u_i < u_j$ or $u_i = u_j$ and $i < j$. In other words, in one-line notation, π_u orders distinct letters of u lexicographically, and equal letters by occurrence order, starting from 0. When u is the BWT image of some necklace, the standard permutation is also called LF-mapping. As is well known, a word u is a BWT image of an aperiodic necklace if and only if π_u is a length-n cycle.

[1] This is the standard convention in the literature. Of course, one can also choose the descending lexicographic order, the properties are symmetric.

The *inverse standard permutation* π_u^{-1} of a word u, written in one-line notation, can be obtained by listing in left-to-right order the positions of 0 in u, then the positions of 1, and so on. The inverse standard permutation of a BWT image is also called FL-mapping.

An aperiodic necklace can be uniquely reconstructed from its BWT, using either of the standard permutation or its inverse. To do this, it is convenient to write these permutations in cycle form. Let L be the BWT of $[w]$, and $\pi_L = (j_0\, j_1 \cdots j_{n-1})$ be the standard permutation of L in cycle form, with $j_0 = 0$. Then, $L_{j_{n-1}} L_{j_{n-2}} \cdots L_{j_0}$ is the first row of the BWT matrix of $[w]$, i.e., the lexicographically least element of the necklace $[w]$. This is also equal to $F_{i_0} F_{i_1} \cdots F_{i_{n-1}}$, where $\pi_L^{-1} = (i_0\, i_1 \cdots i_{n-1})$ is the inverse standard permutation of L in cycle form, with $i_0 = 0$.

The following remark will be used in later sections:

▶ Remark 3. In the case of a balanced Parikh vector $(\frac{n}{k}, \frac{n}{k}, \ldots, \frac{n}{k})$, then, one can reverse the BWT from its inverse standard permutation in cycle form $\pi_u^{-1} = (i_0\, i_1 \cdots i_{n-1})$, $i_0 = 0$, by mapping i_ℓ to $\left\lfloor \dfrac{i_\ell}{n/k} \right\rfloor$ for $\ell = 0, 1, \ldots, n-1$.

For example, the BWT of $[w] = [220120011]$ is the word $u = 202001121$, whose inverse standard permutation in cycle form is $\pi_u^{-1} = (0\,1\,3\,5\,8\,7\,2\,4\,6)$. Mapping: $0, 1, 2$ to 0; $3, 4, 5$ to 1; $6, 7, 8$ to 2, one obtains the word 001122012, the first row of the BWT matrix of $[w]$.

Necklaces that are not aperiodic can also be reconstructed from their BWT, as a consequence of the following result:

▶ **Proposition 4** ([27, Proposition 2]). *For any $c \geq 1$, $u = u_1 u_2 \cdots u_n$ is the BWT of an aperiodic necklace $[w]$ if and only if $u_1^c u_2^c \cdots u_n^c$ is the BWT of $[w^c]$.*

Let $G = (V, E)$ be a finite directed graph, which may have loops and multiple edges (in this case, it is also called a multigraph in the literature). Each edge $e \in E$ is directed from its source vertex $s(e)$ to its target vertex $t(e)$.

The directed *line graph* (or *edge graph*) $\mathcal{L}G = (E, E')$ of G has as vertices the edges of G, and as edges the set

$$E' = \{(e_1, e_2) \in E \times E \mid s(e_2) = t(e_1)\}.$$

An oriented *spanning tree* of G is a subgraph containing all of the vertices of G, having no directed cycles, in which one vertex, the root, has outdegree 0, and every other vertex has outdegree 1. The number $\kappa(G)$ of oriented spanning trees of G is sometimes called the *complexity* of G.

A directed graph is *Hamiltonian* if it has a *Hamiltonian cycle*, i.e., one that traverses each node exactly once, while it is *Eulerian* if it has an *Eulerian cycle*, i.e., one that traverses each edge exactly once. As is well known, a directed graph is Eulerian if and only if $indeg(v) = outdeg(v)$ for all vertices v. If a graph is Eulerian, its line graph is Hamiltonian.

3 Generalized de Bruijn graphs and generalized de Bruijn words

We start by briefly discussing (ordinary) de Bruijn graphs and de Bruijn words, and relating them to the inverse standard permutation of the Burrows–Wheeler transform.

Recall that the de Bruijn graph $\mathrm{DB}(\Sigma_k, k^d)$ of order d is the directed graph whose vertices are the words of length d over Σ_k and there is an edge from $a_i w$ to v, where a_i is a letter, if and only if $v = w a_j$ for some letter a_j. As is well known, de Bruijn graphs are Eulerian and Hamiltonian. One has $\mathcal{L}\,\mathrm{DB}(\Sigma_k, k^d) = \mathrm{DB}(\Sigma_k, k^{d+1})$.

A *de Bruijn word* of order d over Σ_k is a necklace over Σ_k containing as a circular factor each of the Σ_k^d words of length d over Σ_k exactly once. De Bruijn words correspond to Hamiltonian cycles in $\mathrm{DB}(\Sigma_k, k^d)$, as they can be obtained by concatenating the first letter of the labels of the vertices it traverses.

Now, since the vertex labels of $\mathrm{DB}(\Sigma_k, k^d)$ are the base-k representations of the integers in $\{0, 1, \ldots, k^d - 1\}$ on d digits (padded with leading zeroes), the de Bruijn graph of order d over Σ_k can be equivalently defined as the one having vertex set $\{0, 1, \ldots, k^d - 1\}$ and containing an edge from m to $km + i \mod (k^d)$ for every $i = 0, 1, \ldots, k - 1$. We let $\mathrm{DB}(k, k^d)$ denote this equivalent representation of the de Bruijn graph of order d. A Hamiltonian cycle in $\mathrm{DB}(k, k^d)$ is then a cyclic permutation of $\{0, 1, \ldots, k^d - 1\}$. The relation between a de Bruijn word of order d over Σ_k, i.e., a Hamiltonian cycle in $\mathrm{DB}(\Sigma_k, k^d)$, and the corresponding Hamiltonian cycle in $\mathrm{DB}(k, k^d)$, is given in the following lemma.

▶ **Lemma 5.** *Let $[w] = [w_0 \cdots w_{k^d-1}]$ be the necklace encoding a Hamiltonian cycle in $\mathrm{DB}(\Sigma_k, k^d)$, and let $[u] = [u_0 \cdots u_{k^d-1}]$ be the corresponding Hamiltonian cycle in $\mathrm{DB}(k, k^d)$, namely, the one such that w_i is the k-ary expansion on d digits of u_i for each $i = 0, \ldots, k^d - 1$. Then $[u]$ is the inverse standard permutation of the Burrows–Wheeler transform of $[w]$, in cycle form. In particular, one has $w_i = \lfloor u_i / k^{d-1} \rfloor$ for every $0 \leq i \leq k^d - 1$.*

Proof. For a given integer n in $\{0, \ldots, k^d - 1\}$, the first digit of n in its base-k representation on d digits is $\lfloor \frac{n}{k^{d-1}} \rfloor$. The claim then follows directly from Remark 3 and the fact that de Bruijn words have a balanced Parikh vector. ◀

▶ **Example 6.** In $\mathrm{DB}(\Sigma_2, 8)$, the de Bruijn word $[w] = [00010111]$ corresponds to the Hamiltonian cycle obtained by visiting the nodes 000, 001, 010, 101, 011, 111, 110, 100; and to the Hamiltonian cycle $(0, 1, 2, 5, 3, 7, 6, 4)$ in $\mathrm{DB}(2, 8)$. Indeed, one has $\mathrm{BWT}([w]) = 10011010$, and $\pi^{-1}_{\mathrm{BWT}([w])} = (0\,1\,2\,5\,3\,7\,6\,4)$, in cycle notation. By applying the mapping $i \mapsto \lfloor \frac{i}{4} \rfloor$, one retrieves $[w] = [00010111]$.

We now introduce generalized de Bruijn words. In [14], Higgins showed that a word w of length k^n over Σ_k is a (ordinary) de Bruijn word if and only if its BWT is an alphabet-permutation power. This motivates us to introduce the following definition:

▶ **Definition 7.** *A necklace of length kn over Σ_k is a* generalized de Bruijn word *if its BWT is an alphabet-permutation power over Σ_k.*

For example, [02201331] is a generalized de Bruijn word since its BWT is 21302031.

Notice that, in the literature, there already exist several other definitions of generalized de Bruijn words (see, e.g., [4, 1, 28]).

The *generalized de Bruijn graph* $\mathrm{DB}(k, n)$ has vertices $\{0, 1, \ldots, n - 1\}$ and for every vertex m there is an edge from m to $km + i \mod (n)$ for every $i = 0, 1, \ldots, k - 1$. Generalized de Bruijn graphs have been introduced independently by Imase and Itoh [16], and by Reddy, Pradhan, and Kuhl [29] (see also [10]). They are Eulerian and Hamiltonian. By definition, a generalized de Bruijn graph $\mathrm{DB}(k, n)$ is an ordinary de Bruijn graph when $n = k^d$ for some $d > 0$. As an example, the generalized de Bruijn graph $\mathrm{DB}(3, 6)$ is displayed in Fig. 1.

The line graph of $\mathrm{DB}(k, n)$ is $\mathrm{DB}(k, kn)$ [22]. The Eulerian cycles of $\mathrm{DB}(k, n)$ correspond to the Hamiltonian cycles of $\mathrm{DB}(k, kn)$.

In Theorem 9, we give a characterization of generalized de Bruijn words in terms of generalized de Bruijn graphs, which generalizes Lemma 5.

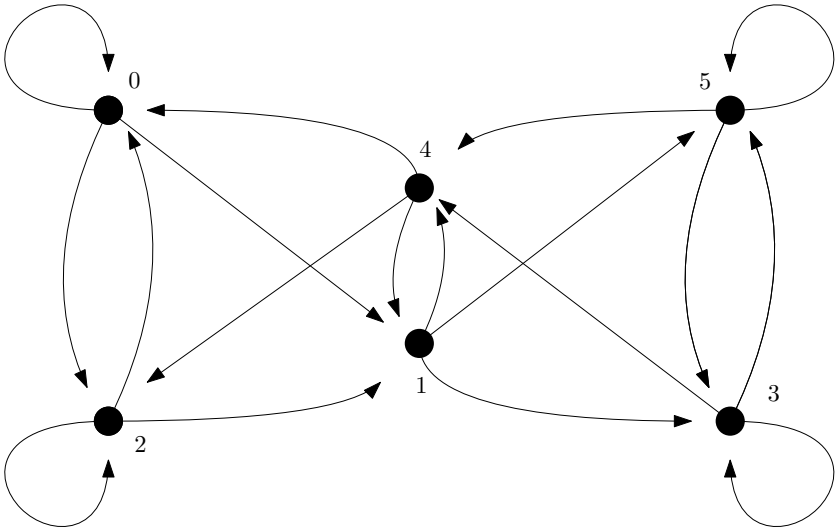

Figure 1 The generalized de Bruijn graph DB(3, 6).

▶ **Lemma 8.** *Let u be a word of length kn over Σ_k, with Parikh vector $(n, n, \dots, n)$. The following statements are equivalent:*

1. *The word u is an alphabet-permutation power;*
2. *For each $0 \le j < k$, $0 \le i < n$, one has $\pi_u^{-1}(i + jn) \in [ki, k(i + 1) - 1]$;*
3. *For each $0 \le \ell < kn$, there is an edge from ℓ to $\pi_u^{-1}(\ell)$ in $\mathrm{DB}(k, kn)$.*

Proof. We first observe that (3) $\iff$ (2). By definition, the edges of the de Bruijn graph $\mathrm{DB}(k, kn)$ are of the form $(\ell, k\ell + i' \bmod kn)$ for $0 \le i' < k$. In particular, for any $\ell = i + jn$ with $0 \le i < n$ and $0 \le j < k$, we have $k(i + jn) + i' \bmod kn \in [ki, k(i + 1) - 1]$. Thus, $\pi_u^{-1}(\ell)$ lies in this interval if and only if there is an edge from ℓ to $\pi_u^{-1}(\ell)$ in $\mathrm{DB}(k, kn)$. Next, we show that (1) implies (2). We start with the following general remark: for a word u and an integer $0 \le \ell \le |u| - 1$, one has $\pi_u^{-1}(\ell) = p$ if p is the position of the ℓth letter in u, according to the order defined by π_u^{-1}, which is the lexicographic order with equal letters sorted by occurrence position. Suppose u is an alphabet-permutation power, so u consists of n consecutive blocks of length k, each of which is a permutation of Σ_k. In particular, each letter appears exactly n times in total. Let $i < n$ and $j < k$. The index $\pi_u^{-1}(i + jn)$ corresponds to the $(i + 1)$th occurrence of the letter j in u. Since u is a concatenation of n blocks, and each block is a permutation of Σ_k, the $(i + 1)$th occurrence of letter j must appear within the $(i + 1)$th block of u, i.e., within positions $[ki, k(i + 1) - 1]$. Therefore, $\pi_u^{-1}(i + jn) \in [ki, k(i + 1) - 1]$, as required. This proves that (1) $\implies$ (2).

It remains to show that (2) implies (1). Assume that for all $0 \le j < k$ and $0 \le i < n$, we have $\pi_u^{-1}(i + jn) \in [ki, k(i + 1) - 1]$. Fix any $i < n$. Then the k values $\pi_u^{-1}(i + jn)$ for $j = 0, \dots, k - 1$ all lie in $[ki, k(i + 1) - 1]$. Since π_u^{-1} is a permutation, these values must be distinct. Therefore, $\{\pi_u^{-1}(i), \pi_u^{-1}(n + i), \dots, \pi_u^{-1}((k - 1)n + i)\} = [ki, k(i + 1) - 1]$.

Let us now examine the content of these positions in u. The value $u[\pi_u^{-1}(i+jn)]$ is the j-th letter in this collection. But $\pi_u^{-1}(i+jn)$ is the position of the $(i+1)$th occurrence of the letter j, hence $u[\pi_u^{-1}(i+jn)] = j$. It follows that $\{u[\pi_u^{-1}(i)], u[\pi_u^{-1}(n+i)], \dots, u[\pi_u^{-1}((k-1)n+i)]\} = \Sigma_k$, which means that the substring $u[ki \,..\, k(i + 1) - 1]$ is a permutation of Σ_k. Since this holds for every $0 \le i < n$, the word u is a concatenation of n permutations of Σ_k, i.e., an alphabet-permutation power. This completes the proof. ◀

As a consequence, we have:

▶ **Theorem 9.** *A necklace is the label of a Hamiltonian cycle of a generalized de Bruijn graph* $\mathrm{DB}(k, kn)$ *if and only if it is the inverse standard permutation, in cycle form, of the BWT of a generalized de Bruijn word of length* kn *over* Σ_k*.*

Proof. The fact that every Hamiltonian cycle of $\mathrm{DB}(k, kn)$ corresponds to the inverse standard permutation of some length n word over k letters follows from condition (2) of Lemma 8 and from the fact that a permutation π is the standard permutation of a word over k letters if and only if there are at most $k-1$ indices i such that $\pi(i+1) < \pi(i)$ [8, Theorem 1]. We use the fact that generalized de Bruijn words are aperiodic necklaces: indeed, if $w = u^r$ is a power (with $r > 1$), then its BWT can be derived from the BWT of u by repeating each letter r times (Proposition 4). Consequently, the BWT of a periodic necklace cannot be an alphabet-permutation power and, conversely, a periodic necklace cannot have a BWT that is an alphabet-permutation power. We now conclude from Lemma 8, and from the fact that a word is a BWT image of an aperiodic necklace if and only if its (inverse) standard permutation is a cycle. ◀

Therefore, the generalized de Bruijn words of length kn over the alphabet Σ_k can be constructed from the Hamiltonian cycles of the generalized de Bruijn graph $\mathrm{DB}(k, kn)$ by mapping $i \mapsto \lfloor i/n \rfloor$ for every $i = 0, \ldots, kn-1$.

▶ **Example 10.** The Hamiltonian cycles in $\mathrm{DB}(3, 6)$ are: $(0, 1, 3, 5, 4, 2)$, $(0, 1, 5, 3, 4, 2)$, $(0, 2, 1, 3, 5, 4)$, and $(0, 2, 1, 5, 3, 4)$. Applying the mapping $i \mapsto \lfloor i/2 \rfloor$, one obtains the ternary generalized de Bruijn words of length 6: [001221], [002121], [010122], and [010212].

Thanks to the characterization given in Theorem 9, we can obtain a formula for counting generalized de Bruijn words. We make use of the BEST theorem for directed graphs, together with the fact that the number of Hamiltonian cycles in $\mathrm{DB}(k, kn)$ is equal to the number of Eulerian cycles in $\mathrm{DB}(k, n)$.

▶ **Theorem 11** (BEST theorem [34])**.** *The number of distinct Eulerian cycles in a Eulerian directed graph* $G = (V, E)$ *is given by*

$$e(G) = \kappa(G) \prod_{j=1}^{n} (d_{v_j} - 1)! \tag{1}$$

where d_{v_j} *is the outdegree of* v_j*.*

By definition, the outdegree of every node in the generalized de Bruijn graph $\mathrm{DB}(k, n)$ is k, therefore $\prod_{j=1}^{n} (d_{v_j} - 1)! = ((k-1)!)^n$. So, we have the following result.

▶ **Theorem 12.** *For every* $n \geq 1$*, the number of generalized de Bruijn words of length* kn *over* Σ_k *is*

$$\mathrm{DBW}_k(kn) = ((k-1)!)^n \kappa(\mathrm{DB}(k, n)). \tag{2}$$

The number of spanning trees can be computed by means of a determinant, as an application of the matrix-tree theorem below. Recall that the *Laplacian matrix* of a directed graph G is the $n \times n$ matrix $L_G = D_G - A_G$, where D_G and A_G are the degree matrix and the adjacency matrix of G, respectively. It can be defined by

$$(L_G)_{ij} = \begin{cases} d_{v_i} - d_{v_i v_i} & \text{if } i = j \\ -d_{v_i v_j} & \text{if } i \neq j \end{cases}$$

where $d_{v_i v_j}$ is the number of edges directed from v_i to v_j, and $d_{v_i} = \sum_j d_{v_i v_j}$ is the outdegree of v_i, so that $d_{v_i} - d_{v_i v_i}$ is the outdegree of v_i minus the number of its self loops.

Table 1 The first few values of $\mathrm{DBW}_2(2n)$ and $\mathrm{DBW}_3(3n)$ (resp. sequence A027362 and A192513 in [32]).

n	2	3	4	5	6	7	8	9	10	11
$\mathrm{DBW}_2(2n)$	1	1	2	3	4	7	16	21	48	93
$\mathrm{DBW}_3(3n)$	4	24	64	512	1728	13312	32768	373248	1310720	10903552

▶ **Theorem 13** (Matrix-Tree Theorem for Eulerian graphs [17]). *Let G be an Eulerian graph. The number $\kappa(G)$ is equal to the determinant of the matrix obtained from the Laplacian matrix of G after removing the last row and the last column.*

The matrix L_G^0 obtained by removing the last row and column from L_G is sometimes called the *reduced Laplacian matrix* of G.

We now discuss the relations between generalized de Bruijn words and *sandpile groups*, algebraic objects capturing important properties of directed Eulerian graphs, which can be defined in terms of the reduced Laplacian matrix of the graph. We start by giving some standard results and definitions from [33]. Recall that any integer matrix A can be written in a canonical form $A = PDQ$, where both matrices P, Q are integer matrices and D is an integer diagonal matrix with the property that if $d_1, \dots, d_n$ are the nonzero entries of the main diagonal of D, then $d_i | d_{i+1}$ for every i. The matrix D is called the *Smith Normal Form* of A. The d_i are called the *invariant factors* of the matrix A.

Let $G = (V, E)$ be a finite strongly connected directed graph, that is, for any $v, w \in V$ there are directed paths in G from v to w and from w to v. Then associated to any vertex v of G is an Abelian group $K(G, v)$, the *sandpile group* (also known as critical group, Picard group, Jacobian, and group of components, see [33] for the detailed definition). When G is Eulerian, the groups $K(G, v)$ and $K(G, u)$ are isomorphic for any $v, u \in V$, and we let $K(G)$ denote the *sandpile group of* G, and one has:

$$K(G) \cong \bigoplus_{i=1}^{n-1} \mathbb{Z}_{d_i} \tag{3}$$

where $d_1, \dots, d_{n-1}$ are the invariant factors of L_G^0, or equivalently, $d_1, \dots, d_{n-1}, 0$ are the invariant factors of L_G.

By the matrix-tree theorem, the order of the sandpile group $|K(G)| = \kappa(G)$ is equal to $\det(L_G^0)$, the determinant of the reduced Laplacian matrix of G.

▶ **Example 14.** Consider the generalized de Bruijn graph $\mathrm{DB}(3, 6)$ (Fig. 1). Its Laplacian matrix is

$$L_{\mathrm{DB}(3,6)} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ 0 & 3 & 0 & -1 & -1 & -1 \\ -1 & -1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & -1 & -1 \\ -1 & -1 & -1 & 0 & 3 & 0 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

and its Smith Normal Form is

$$\mathrm{SNF}(L_{\mathrm{DB}(3,6)}) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Hence, by (3), we have $K(\mathrm{DB}(3,6)) \cong \mathbb{Z}_3^3$. The matrix $\mathrm{SNF}(L_{\mathrm{DB}(3,6)})$ has determinant $3^3 = 27$, so $\kappa(\mathrm{DB}(3,6)) = 27$. Thus, by Theorem 12, there are $27 \cdot 2^6 = 1728$ ternary generalized de Bruijn words of length 18 – the same as the number of distinct Eulerian cycles in $\mathrm{DB}(3,6)$ and Hamiltonian cycles in $\mathrm{DB}(3,18)$.

As another example, the Laplacian matrix of $\mathrm{DB}(2,6)$ is

$$L_{\mathrm{DB}(2,6)} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 2 & -1 & -1 & 0 & 0 \\ 0 & 0 & 2 & 0 & -1 & -1 \\ -1 & -1 & 0 & 2 & 0 & 0 \\ 0 & 0 & -1 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

whose Smith Normal Form is

$$\mathrm{SNF}(L_{\mathrm{DB}(2,6)}) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

We have $K(\mathrm{DB}(2,6)) \cong \mathbb{Z}_2^2$, and in fact there are $4 \cdot (1!)^6 = 4$ binary generalized de Bruijn words of length 12, namely [000010111101], [000011101101], [000100101111], and [000100111011].

For ordinary de Bruijn graphs over a binary alphabet, the general structure of the sandpile group was described by Levine [20]:

$$K(\mathrm{DB}(2, 2^d)) \cong \bigoplus_{i=1}^{d-1} (\mathbb{Z}_{2^i})^{2^{d-1-i}}. \tag{4}$$

In the case of generalized de Bruijn graphs, Chan, Hollmann and Pasechnik gave the structure of $K(\mathrm{DB}(k,n))$ for arbitrary k and n [5, Theorem 3.1]. We refer the reader to [5, 6] for further details.

In the next sections, we focus on the case where the alphabet size is a prime number p, allowing us to connect generalized de Bruijn words to necklaces having a BWT matrix that is invertible over $\mathbb{Z}_p$.

4 Invertible necklaces and Reutenauer groups

In this section, we define invertible necklaces and highlight their connections with abstract algebra.

In [35] (see also the related paper [30]), the authors used BWT matrix multiplication to obtain a group structure on important classes of binary words (namely, *Christoffel words* and *balanced* words) that have an invertible BWT matrix. Following this idea, we define *invertible* necklaces:

▶ **Definition 15.** *Let p be a prime. A necklace $[w]$ over the alphabet $\Sigma_p = \{0, 1, \ldots, p-1\}$ is called* invertible *if its BWT matrix is non-singular, i.e., has nonzero determinant modulo p.*

As observed in [35], the product of two BWT matrices is not, in general, a BWT matrix; but the author suggests replacing BWT matrices by *circulant matrices*. As we will see, using circulant matrices, we can define the product between any two invertible necklaces.

Let w be a word over Σ_k. The *circulant matrix* of w is the square matrix $\mathcal{C}_w$ whose $(i+1)$th row is $\sigma^i(w)$. For example, the circulant matrix of 011 is $\mathcal{C}_{011} = \begin{pmatrix} 0\,1\,1 \\ 1\,0\,1 \\ 1\,1\,0 \end{pmatrix}$.

One can immediately observe that, since the circulant matrix of a word can be obtained by permuting the rows of its BWT matrix, a necklace is invertible if and only if any (and then, every) of its associated circulant matrices is invertible.

Recall that for every prime p and any positive integer n, there is a unique finite field with p^n elements, denoted $\mathbb{F}_{p^n}$, and its multiplicative group $\mathbb{F}_{p^n}^*$ is cyclic. The set of invertible $n \times n$-circulant matrices over $\mathbb{F}_p$ forms, with respect to matrix multiplication, an Abelian group $C(p,n)$. We consider the group $C(p,n)/\langle Q_n \rangle$, where Q_n is the permutation matrix of the cycle $(0\,1\,\ldots\,n-1)$. This group is called the *Reutenauer group* RG_p^n in [11]. Intuitively, an element of the Reutenauer group corresponds to an equivalence class of invertible circulant matrices, where two circulant matrices are equivalent if and only if they are the circulant matrices of two conjugated words. Thus, RG_p^n is in natural bijection with the set of invertible necklaces of length n over Σ_p, and one can write $\mathcal{C}_{[w]} = \{\mathcal{C}_v,\ v \in [w]\} \in RG_p^n$ for the class of circulant matrices associated with a given invertible necklace $[w]$ of length n over Σ_p. In particular, this induces a group structure on the set of such invertible necklaces, isomorphic to RG_p^n, namely with respect to the multiplication defined by $[u] \cdot [v] = [w]$ where $[w]$ is such that $\mathcal{C}_{[u]} \cdot \mathcal{C}_{[v]} = \mathcal{C}_{[w]}$. This multiplication does not depend on the choice of representatives for each necklace (equivalently, for each matrix). The *trace* of a matrix is defined as the sum of its diagonal coefficients. For a word w, one can observe that $\mathcal{C}_w$ has trace $wt(w)$. Since $wt(w)$ is invariant under rotation, one can define the trace of a class of circulant matrices $\mathcal{C}_{[w]}$ as $\operatorname{Tr} \mathcal{C}_{[w]} = wt([w]) \mod p$.

As shown in [11], the Reutenauer group acts on the set of all aperiodic necklaces as follows: Let $[v]$ be an aperiodic necklace of length n over Σ_p. For any $\mathcal{C}_{[w]} \in RG_p^n$, we define $\mathcal{C}_{[w]} \cdot [v]$ as the necklace of $\mathcal{C}_w \cdot v^T$. In particular, this operation does not depend on the choice of a representative. For example, $\mathcal{C}_{[11010]} \cdot [11000] = [11110]$ since $\begin{pmatrix} 1\,1\,0\,1\,0 \\ 0\,1\,1\,0\,1 \\ 1\,0\,1\,1\,0 \\ 0\,1\,0\,1\,1 \\ 1\,0\,1\,0\,1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = 01111$.

The orbit of an aperiodic necklace $[v]$ is therefore $O_{[v]} = \{[\mathcal{C}_{[w]} \cdot [v]] \mid \mathcal{C}_{[w]} \in RG_p^n\}$.

We now recall some classical results and definitions in abstract algebra (see, for example, [23] for a more detailed presentation), which have a strong connection with circulant matrices and invertible necklaces.

It is well known that the number of irreducible polynomials over $\mathbb{F}_p$ of degree n is equal to the number of aperiodic necklaces of length n over Σ_p, but there is no known canonical bijection between the two sets. The *minimal polynomial* of a nonzero element α of $\mathbb{F}_{p^n}$ is the (unique) monic polynomial $f \in \mathbb{F}_p[X]$ with the least degree for which α is a root. This polynomial has as roots all the m distinct elements of the form $\alpha, \alpha^p, \ldots, \alpha^{p^{m-1}}$, where m is the least integer such that $\alpha^{p^m} = \alpha$, i.e., the distinct roots of f are the orbit of α under the application of the Frobenius automorphism $\alpha \mapsto \alpha^p$. Thus, f has degree m and can be

factored as

$$f = (x-\alpha)(x-\alpha^p)\cdots(x-\alpha^{p^{m-1}}).$$

The sum of all roots of f is called the *trace* of f.

But $\mathbb{F}_{p^n}$ is also a vector space over $\mathbb{F}_p$. An element γ of $\mathbb{F}_{p^n}$ is called *normal* if $\{\gamma, \gamma^p, \ldots, \gamma^{p^{n-1}}\}$ forms a vector space basis for $\mathbb{F}_{p^n}$ over $\mathbb{F}_p$, called a *normal basis*.

The minimal polynomial of a normal element of $\mathbb{F}_{p^n}$ over $\mathbb{F}_p$ is called a *normal polynomial*. Normal polynomials have non-zero trace (modulo p).

If γ is a normal element of $\mathbb{F}_{p^n}$, every element of $\mathbb{F}_{p^n}$ can be written as a linear combination of elements in the normal basis $\{\gamma, \gamma^p, \ldots, \gamma^{p^{n-1}}\}$. That is, we can write all the elements of $\mathbb{F}_{p^n}$ as n-length words, or n-length vectors. More precisely, if $\alpha = w_0\gamma + w_1\gamma^p + \ldots + w_{n-1}\gamma^{p^{n-1}}$, $w_j \in \mathbb{F}_p$ for every $j = 0, \ldots, n-1$, we can represent α with the word $w = w_0 w_1 \cdots w_{n-1}$. The application of the Frobenius automorphism to an element of the field corresponds to the shift of the corresponding word, and an orbit to a conjugacy class, i.e., to a necklace. This necklace is aperiodic if and only if the orbit is maximal, i.e., has size n.

Table 2 The finite field $\mathbb{F}_{2^4}$ as a vector space over $\mathbb{F}_2$, where the elements are grouped by orbits with respect to a normal element γ.

orbit	orbit as vectors (words)	normal
$\{0\}$	$\{0000\}$	no
$\{\gamma, \gamma^2, \gamma^4, \gamma^8\}$	$\{1000, 0100, 0010, 0001\}$	yes
$\{\gamma+\gamma^2, \gamma^2+\gamma^4, \gamma^4+\gamma^8, \gamma+\gamma^8\}$	$\{1100, 0110, 0011, 1001\}$	no
$\{\gamma^2+\gamma^8, \gamma+\gamma^4\}$	$\{0101, 1010\}$	no
$\{\gamma+\gamma^2+\gamma^4, \gamma^2+\gamma^4+\gamma^8, \gamma+\gamma^4+\gamma^8, \gamma+\gamma^2+\gamma^8\}$	$\{1110, 0111, 1011, 1101\}$	yes
$\{\gamma+\gamma^2+\gamma^4+\gamma^8\}$	$\{1111\}$	no

Fixing a normal element γ and writing elements of $\mathbb{F}_{p^n}$ in the induced normal basis, the orbit of γ corresponds to the necklace $[10^{n-1}]$. There may be other orbits constituted by normal elements. For example, the orbit corresponding to the necklace $[1110]$ is another orbit of normal elements in $\mathbb{F}_{2^4}$.

The number of normal elements of $\mathbb{F}_{p^n}$ is counted by $\Phi_p(n)$, the generalized Euler's totient function. Equivalently, $\Phi_p(n)$ counts the number of polynomials over $\mathbb{F}_p$ of degree smaller than n and coprime with $X^n - 1$. It can be computed using the formula

$$\Phi_p(n) = p^n \prod_{d \mid (n/\lambda_p(n))} \left(1 - \frac{1}{p^{\mathrm{ord}_p(d)}}\right)^{\frac{\phi(d)}{\mathrm{ord}_p(d)}}$$

for $n > 1$, where $\lambda_p(n)$ is the largest power of p dividing n, and $\mathrm{ord}_p(d)$ is the multiplicative order of d modulo p. The first mention of this formula seems to come from [13], and a more recent study can be found in [15].

The first few values of the sequences $\Phi_2(n)$ and $\Phi_3(n)$ are presented in Table 3.

Table 3 First few values of $\Phi_2(n)$ and $\Phi_3(n)$ (resp. sequence A003473 and A003474 in [32]).

n	1	2	3	4	5	6	7	8	9	10	11	12
$\Phi_2(n)$	1	2	3	8	15	24	49	128	189	480	1023	1536
$\Phi_3(n)$	1	4	18	32	160	324	1456	2048	13122	25600	117128	209952

Let p be a prime, and consider words in Σ_p^n as vectors in the vector space $\mathbb{F}_{p^n}$.

An $n \times n$-circulant matrix over $\mathbb{F}_p$ is in $C(p,n)$ if and only if it is invertible (or non-singular), if and only if its determinant is non-zero modulo p; or, equivalently, if its rows form a normal basis of $\mathbb{F}_{p^n}$.

Since every element of a normal basis can be chosen as the first row of a corresponding invertible circulant matrix, the order of $C(p,n)$ is $\Phi_p(n)$, and the order of RG_p^n (hence, the number of invertible necklaces) is $\Phi_p(n)/n$.

We therefore arrive at the following characterization:

▶ **Theorem 16.** *Let $[w]$ be a necklace over Σ_p, p prime. The following are equivalent:*

1. *The necklace $[w]$ is invertible;*
2. *Any word in $[w]$ has an invertible circulant matrix over $\mathbb{F}_p$;*
3. *$\mathcal{C}_{[w]}$ belongs to the Reutenauer group RG_p^n;*
4. *Any word in $[w]$ is a vector corresponding to a normal element of $\mathbb{F}_{p^n}$;*
5. *Every word of length n over Σ_p can be written in a unique way as a sum (modulo p) of words in $[w]$.*

▶ **Example 17.** Let $p = 2$ and $n = 4$. We have three aperiodic binary necklaces, namely $[1000]$, $[1100]$, and $[1110]$. Out of them, the first and the last are invertible, while the second is not, since $\det \begin{pmatrix} 1&1&0&0 \\ 0&1&1&0 \\ 0&0&1&1 \\ 1&0&0&1 \end{pmatrix} = 0$. Indeed, $1100 + 0110 + 0011 = 1001$ in $\mathbb{F}_{2^4}$, so the elements are not linearly independent as vectors, and therefore do not form a basis of $\mathbb{F}_{2^4}$.

The Reutenauer group RG_2^4 is therefore constituted by the classes of matrices $Id = \mathcal{C}_{[1000]}$ and $\mathcal{C}_{[1110]}$.

So, the set of aperiodic necklaces of length n over Σ_p with invertible BWT matrix forms an abelian group isomorphic to the Reutenauer group RG_p^n, but the multiplication has to be carried out with the circulant matrices rather than with BWT matrices. This was also speculated in a remark we recently found at the end of [35] (see also the related paper [30]).

▶ **Lemma 18.** *Let $[w]$ be a necklace over Σ_p and let f be the minimal polynomial of w, seen as a vector. Then, f has trace $(m, \dots, m)$, where $m = wt([w]) \mod p$. If $[w]$ is invertible, then $m = \operatorname{Tr} \mathcal{C}_{[w]} \neq 0$.*

Proof. Since the roots of f are, as vectors, shifts of w, the trace of f is the vector $(m, \dots, m)$. Furthermore, if w is normal, all elements of $\mathcal{C}_{[w]}$ are invertible, and the trace $\operatorname{Tr} \mathcal{C}_{[w]}$ must be nonzero. ◀

In particular, binary invertible necklaces have an odd number of 1's. However, note that, in general, having a non-zero trace is not a sufficient condition for an element to be normal.

Recall that a prime q is called *p-rooted* if p is a generator of the cyclic group $\mathbb{F}_q^*$. For example, 5 is 2-rooted since $\{2^i \mod 5 \mid i = 1, \dots, 4\} = \mathbb{F}_5^*$, while 7 is not 2-rooted since $\{2^i \mod 7 \mid i = 1, \dots, 6\} = \{1, 2, 4\}$.

In [7], the authors obtained the following result:

▶ **Theorem 19** ([7])**.** *Let n be a positive integer. Every monic irreducible polynomial of degree n over $\mathbb{F}_p$ with nonzero trace is normal if and only if n is either a power of p or a p-rooted prime.*

The latter result, together with Lemma 18, immediately implies the following:

▶ **Corollary 20.** *Let n be a positive integer. Every aperiodic necklace of length n over Σ_p with non-zero weight modulo p is invertible if and only if n is either a power of p or a p-rooted prime.*

For example, the sequence of 2-rooted primes is A001122 in OEIS [32]. Therefore, by Corollary 20, this sequence precisely corresponds to the lengths n (excluding powers of 2) for which all binary necklaces of non-zero weight modulo 2 are invertible. This latter is also the sequence of numbers of the form $2m+1$ such that $(10)^m$ is a BWT image (see [2, 26]), i.e., lengths for which one can have a BWT with the largest possible number of runs (which could be considered as the *worst case* of the BWT). Notice that words having BWT of the form $(10)^m$ are, by definition, generalized de Bruijn words.

However, we do not know whether there exist infinitely many n, different from a power of p, for which every aperiodic necklace of length n over Σ_p with non-zero weight modulo p is invertible. This seems to be a difficult problem, since it is related to the famous conjecture of Emil Artin stating that a given integer a that is neither a square number nor -1 is a primitive root modulo infinitely many primes p.

5 Generalized de Bruijn words and invertible necklaces

In this section, we further explore the connection between the content of Sections 3 and 4. Specifically, we show that over an alphabet whose size is a prime number, generalized de Bruijn words and invertible necklaces are linked not only through their relationship with the Burrows–Wheeler transform, but also via an isomorphism of abelian groups.

First, recall that every finite abelian group G is isomorphic to the direct sum of cyclic groups, i.e., $G \cong \bigoplus_i \mathbb{Z}_{d_i}$. One has, for example, $RG_2^4 \cong \mathbb{Z}_2$ and $RG_2^8 \cong \mathbb{Z}_2^2 \oplus \mathbb{Z}_4$ (remember that the order of RG_p^n is $\Phi_p(n)/n$).

Chan, Hollmann and Pasechnik [5] proved that the structure of Reutenauer groups can be found by means of an isomorphism with the sandpile groups of generalized de Bruijn graphs, as reported in the next theorem.

▶ **Theorem 21** ([5])**.** *Let $K(\mathrm{DB}(p,n))$ be the sandpile group of the generalized de Bruijn graph $\mathrm{DB}(p,n)$, p prime. Then*

$$K(\mathrm{DB}(p,n)) \oplus \mathbb{Z}_{p-1} \cong RG_p^n,$$

where $RG_p^n \cong C(p,n)/\langle Q_n \rangle$ is the n-th Reutenauer group over $\mathbb{F}_p$.

Thus, for p prime, $K(\mathrm{DB}(p,n)) \cong C(p,n)/(\mathbb{Z}_{p-1} \times \mathbb{Z}_n)$, and therefore $\kappa(DB(p,n)) = \dfrac{\Phi_p(n)}{n(p-1)}$.

As a consequence of Theorem 21, we arrive at a remarkable fact: the structure of the Reutenauer groups is entirely determined by the structure of the generalized de Bruijn graphs. An illustrative example is provided in Table 4. We point out that the (reduced) Laplacian matrix of a generalized de Bruijn graph (as well as its Smith Normal Form and invariant factors) can be derived directly from the arithmetic definition of the graph, without the need to construct it explicitly.

We thus obtain the following formula for the number of generalized de Bruijn words over an alphabet of prime size:

Table 4 The invariant factors of the reduced Laplacian matrix of the generalized de Bruijn graphs for $k = 3$, which give the structure of the sandpile groups $K(\mathrm{DB}(3, n))$ and Reutenauer groups RG_3^n.

n	d_i	$K(\mathrm{DB}(3,n))$	RG_3^n
3	$(1, 3)$	$\mathbb{Z}_3$	$\mathbb{Z}_3 \oplus \mathbb{Z}_2$
4	$(1, 1, 4)$	$\mathbb{Z}_4$	$\mathbb{Z}_4 \oplus \mathbb{Z}_2$
5	$(1, 1, 1, 16)$	$\mathbb{Z}_{16}$	$\mathbb{Z}_{16} \oplus \mathbb{Z}_2$
6	$(1, 1, 3, 3, 3)$	$\mathbb{Z}_3^3$	$\mathbb{Z}_3^3 \oplus \mathbb{Z}_2$
7	$(1, 1, 1, 1, 1, 104)$	$\mathbb{Z}_{104}$	$\mathbb{Z}_{104} \oplus \mathbb{Z}_2$
8	$(1, 1, 1, 1, 2, 8, 8)$	$\mathbb{Z}_8^2 \oplus \mathbb{Z}_2$	$\mathbb{Z}_8^2 \oplus \mathbb{Z}_2^2$
9	$(1, 1, 1, 3, 3, 3, 3, 9)$	$\mathbb{Z}_9 \oplus \mathbb{Z}_3^4$	$\mathbb{Z}_9 \oplus \mathbb{Z}_3^4 \oplus \mathbb{Z}_2$
10	$(1, 1, 1, 1, 1, 1, 1, 16, 80)$	$\mathbb{Z}_{80} \oplus \mathbb{Z}_{16}$	$\mathbb{Z}_{80} \oplus \mathbb{Z}_{16} \oplus \mathbb{Z}_2$
11	$(1, 1, 1, 1, 1, 1, 1, 1, 22, 242)$	$\mathbb{Z}_{242} \oplus \mathbb{Z}_{22}$	$\mathbb{Z}_{242} \oplus \mathbb{Z}_{22} \oplus \mathbb{Z}_2$
12	$(1, 1, 1, 1, 3, 3, 3, 3, 3, 3, 12)$	$\mathbb{Z}_{12} \oplus \mathbb{Z}_3^6$	$\mathbb{Z}_{12} \oplus \mathbb{Z}_3^6 \oplus \mathbb{Z}_2$

▶ **Theorem 22.** *Let p be a prime. For every $n \geq 2$, the number of generalized de Bruijn words of length pn over $\Sigma_p = \{0, 1, \ldots, p-1\}$ is*

$$\mathrm{DBW}_p(pn) = \frac{((p-1)!)^n}{p-1}\frac{\Phi_p(n)}{n} = \frac{((p-1)!)^n}{p-1}\hat{\mathrm{N}}_p(n) \tag{5}$$

where $\hat{\mathrm{N}}_p(n)$ is the number of invertible necklaces of length n over Σ_p.

Proof. By Theorem 21, we have $\kappa(\mathrm{DB}(p, n)) = \Phi_p(n)/(n(p-1))$. The result then follows from (2). ◀

▶ Remark 23. When n is a power of a prime p (which corresponds to the case of ordinary de Bruijn words), the precise structure of the sandpile group (and hence of the Reutenauer group) can be easily given in terms of p and n. Indeed, from the results in [5], one can derive:

$$K(\mathrm{DB}(p, p^n)) \cong \Big[\bigoplus_{i=1}^{n-1}(\mathbb{Z}_{p^i})^{p^{n-1-i}(p^2-2p+1)}\Big] \oplus \mathbb{Z}_{p^n}^{p-2}, \tag{6}$$

and hence, by Theorem 21,

$$RG_p^n \cong \Big[\bigoplus_{i=1}^{n-1}(\mathbb{Z}_{p^i})^{p^{n-1-i}(p^2-2p+1)}\Big] \oplus \mathbb{Z}_{p^n}^{p-2} \oplus \mathbb{Z}_{p-1}. \tag{7}$$

Note that, as expected, for $p = 2$ we obtain (4).

▶ Remark 24. Let $pn = p^d$, $d > 1$, i.e., $n = p^{d-1}$. Since $\Phi_p(p^{d-1}) = p^{p^{d-1}}\dfrac{p-1}{p}$, from (5) we get

$$\mathrm{DBW}_p(p^d) = \frac{((p-1)!)^{p^{d-1}}p^{p^{d-1}}}{pp^{d-1}} = \frac{(p!)^{p^{d-1}}}{p^d}$$

i.e., the number of (ordinary) de Bruijn words of order d over Σ_p.

In the special case $p = 2$, we have, by Theorem 21, that $K(\mathrm{DB}(2, n)) \cong RG_2^n$. As a consequence, we have the following:

▶ **Theorem 25.** *For every $n \geq 1$, the following sets are in bijection:*

- *The set of binary generalized de Bruijn words of length $2n$;*
- *The set of binary invertible necklaces of length n;*
- *The set of (unordered) normal bases of $\mathbb{F}_{2^n}$.*

In particular, when $n = 2^d$ for some d, we have a bijection between the set of binary de Bruijn words of length 2^{d+1} and the set of binary necklaces of length 2^d having an odd number of 1's. This follows from Corollary 20. Indeed, notice that we do not need to add the hypothesis that the necklaces are aperiodic, since their lengths are a power of 2.

▶ **Theorem 26.** *For every $d \geq 1$, the following sets are in bijection:*

- *The set of binary de Bruijn words of order $d + 1$;*
- *The set of binary necklaces of length 2^d having an odd number of 1's;*
- *The set of (unordered) normal bases of $\mathbb{F}_{2^{2^d}}$.*

▶ Remark 27. Any constructive bijection would allow one to derive algorithms for interchanging between these objects. One could, for example, use such a bijection to generate every binary de Bruijn word with uniform probability, a problem that is still open – in [24] the authors presented an efficient algorithm to generate every binary de Bruijn word with positive probability. Moreover, this could be made efficient in practice since the sum in $\mathbb{F}_2$ is nothing else than the XOR.

6 Conclusions and Open Problems

As emphasized in the introduction, the central motivation of our work is to present a new formalism, based on combinatorics on words, that bridges various algebraic structures. This approach builds on known results relating sandpile groups of (generalized) de Bruijn graphs and groups of invertible matrices over finite fields (Reutenauer groups). The new classes of words we introduce in this paper – generalized de Bruijn words on the one hand, and conjugacy classes of words (necklaces) with invertible Burrows–Wheeler matrix on the other hand – offer a fresh reinterpretation of these connections in terms of words. The key advantage of the new framework is its foundation in the Burrows–Wheeler transform, which not only clarifies links to longstanding open problems in elementary number theory but also opens new avenues for further research.

We expect that different research communities, including those focused on combinatorics on words, commutative algebra, combinatorial algorithms, finite fields, and cryptography, will build upon the new constructions introduced in this paper. For instance, it would be interesting to develop combinatorial characterizations of the newly introduced classes of words. Another compelling direction is to gain a deeper understanding of the group operation on invertible necklaces and investigate its relationship to the one recently introduced by Zamboni in the context of Christoffel words [35], and further explored in [30].

References

1 Yu Hin Au. Generalized de Bruijn words for primitive words and powers. *Discret. Math.*, 338(12):2320–2331, 2015. doi:10.1016/J.DISC.2015.05.025.

2 D. J. Aulicino and Morris Goldfeld. A new relation between primitive roots and permutations. *The American Mathematical Monthly*, 76(6):664–666, 1969. doi:10.1080/00029890.1969.12000297.

3 Jean Berstel, Aaron Lauve, Christophe Reutenauer, and Franco Saliola. *Combinatorics on Words: Christoffel Words and Repetition in Words*, volume 27 of *CRM monograph series*. American Mathematical Society, 2008.

4 Lenore Blum, Manuel Blum, and Michael Shub. Comparison of Two Pseudo-Random Number Generators. In David Chaum, Ronald L. Rivest, and Alan T. Sherman, editors, *Advances in Cryptology*, pages 61–78, Boston, MA, 1983. Springer US.

5 Swee Hong Chan, Henk D. L. Hollmann, and Dmitrii V. Pasechnik. Critical groups of generalized de Bruijn and Kautz graphs and circulant matrices over finite fields. In Jaroslav Nešetřil and Marco Pellegrini, editors, *The Seventh European Conference on Combinatorics, Graph Theory and Applications*, pages 97–104, Pisa, 2013. Scuola Normale Superiore.

6 Swee Hong Chan, Henk D.L. Hollmann, and Dmitrii V. Pasechnik. Sandpile groups of generalized de Bruijn and Kautz graphs and circulant matrices over finite fields. *Journal of Algebra*, 421:268–295, 2015. Special issue in memory of Ákos Seress. doi:10.1016/j.jalgebra.2014.08.029.

7 Yaotsu Chang, T. K. Truong, and I. S. Reed. Normal bases over $GF(q)$. *Journal of Algebra*, 241:89–101, 2001.

8 Maxime Crochemore, Jacques Désarménien, and Dominique Perrin. A note on the burrows - wheeler transformation. *Theor. Comput. Sci.*, 332(1-3):567–572, 2005. doi:10.1016/J.TCS.2004.11.014.

9 Maxime Crochemore, Christophe Hancart, and Thierry Lecroq. *Algorithms on strings*. Cambridge University Press, 2007.

10 Ding-Zhu Du and Frank K. Hwang. Generalized de Bruijn digraphs. *Networks*, 18(1):27–38, 1988. doi:10.1002/NET.3230180105.

11 S. V. Duzhin and D. V. Pasechnik. Groups acting on necklaces and sandpile groups. *Journal of Mathematical Sciences*, 200(6):690–697, 2014. doi:10.1007/s10958-014-1960-6.

12 Gabriele Fici, Sabrina Mantaci, Antonio Restivo, Giuseppe Romana, Giovanna Rosone, and Marinella Sciortino. BWT and Combinatorics on Words. In Paolo Ferragina, Travis Gagie, and Gonzalo Navarro, editors, *The Expanding World of Compressed Data: A Festschrift for Giovanni Manzini's 60th Birthday*, volume 131 of *OASIcs*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025.

13 K. Hensel. Ueber die darstellung der zahlen eines gattungsbereiches für einen beliebigen primdivisor. *Journal für die reine und angewandte Mathematik*, 1888(103):230–237, 1888. doi:10.1515/crll.1888.103.230.

14 Peter M. Higgins. Burrows-Wheeler transformations and de Bruijn words. *Theor. Comput. Sci.*, 457:128–136, 2012. doi:10.1016/J.TCS.2012.07.019.

15 Trevor Hyde. Normal elements in finite fields, 2018. arXiv:1809.02155.

16 Makoto Imase and Masaki Itoh. Design to minimize diameter on building-block network. *IEEE Trans. Computers*, 30(6):439–442, 1981. doi:10.1109/TC.1981.1675809.

17 G. Kirchhoff. On the solution of the equations obtained from the investigation of the linear distribution of galvanic currents. *IRE Transactions on Circuit Theory*, 5(1):4–7, 1958. doi:10.1109/TCT.1958.1086426.

18 Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, 2012.

19 Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009.

20 Lionel Levine. Sandpile groups and spanning trees of directed line graphs. *J. Comb. Theory A*, 118(2):350–364, 2011. doi:10.1016/J.JCTA.2010.04.001.

21 Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010. doi:10.1093/BIOINFORMATICS/BTP698.

22 Xueliang Li and Fuji Zhang. On the numbers of spanning trees and Eulerian tours in generalized de Bruijn graphs. *Discret. Math.*, 94(3):189–197, 1991. doi:10.1016/0012-365X(91)90024-V.

23 Rudolf Lidl and Harald Niederreiter. *Finite Fields*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2 edition, 1996.

24 Zsuzsanna Lipták and Luca Parmigiani. A BWT-Based Algorithm for Random de Bruijn Sequence Construction. In José A. Soto and Andreas Wiese, editors, *LATIN 2024: Theoretical Informatics - 16th Latin American Symposium, Puerto Varas, Chile, March 18-22, 2024, Proceedings, Part I*, volume 14578 of *Lecture Notes in Computer Science*, pages 130–145. Springer, 2024. doi:10.1007/978-3-031-55598-5_9.

25 M. Lothaire. *Combinatorics on Words*. Advanced Book Program. Addison-Wesley, Advanced Book Program, World Science Division, 1983.

26 Sabrina Mantaci, Antonio Restivo, Giovanna Rosone, Marinella Sciortino, and Luca Versari. Measuring the clustering effect of BWT via RLE. *Theor. Comput. Sci.*, 698:79–87, 2017. doi:10.1016/J.TCS.2017.07.015.

27 Sabrina Mantaci, Antonio Restivo, and Marinella Sciortino. Burrows-Wheeler transform and Sturmian words. *Inf. Process. Lett.*, 86(5):241–246, 2003. doi:10.1016/S0020-0190(02)00512-4.

28 Abhinav Nellore and Rachel A. Ward. Arbitrary-Length Analogs to de Bruijn Sequences. In Hideo Bannai and Jan Holub, editors, *33rd Annual Symposium on Combinatorial Pattern Matching, CPM 2022, June 27-29, 2022, Prague, Czech Republic*, volume 223 of *LIPIcs*, pages 9:1–9:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022. doi:10.4230/LIPICS.CPM.2022.9.

29 S.M. Reddy, D.K. Pradhan, and J.G. Kuhl. Directed graphs with minimum diameter and maximal connectivity. Technical report, School of Engineering, Oakland University, July 1980.

30 Christophe Reutenauer and Jeffrey O. Shallit. Christoffel Matrices and Sturmian Determinants. *CoRR*, abs/2409.09824, 2024. doi:10.48550/arXiv.2409.09824.

31 Giovanna Rosone and Marinella Sciortino. The Burrows-Wheeler Transform between Data Compression and Combinatorics on Words. In Paola Bonizzoni, Vasco Brattka, and Benedikt Löwe, editors, *The Nature of Computation. Logic, Algorithms, Applications - 9th Conference on Computability in Europe, CiE 2013, Milan, Italy, July 1-5, 2013. Proceedings*, volume 7921 of *Lecture Notes in Computer Science*, pages 353–364. Springer, 2013. doi:10.1007/978-3-642-39053-1_42.

32 N. J. A. Sloane. The On-Line Encyclopedia of Integer Sequences. Available electronically at http://oeis.org.

33 Richard P. Stanley. Smith normal form in combinatorics. *J. Comb. Theory A*, 144:476–495, 2016. doi:10.1016/J.JCTA.2016.06.013.

34 T. van Aardenne-Ehrenfest and N.G. de Bruijn. *Circuits and Trees in Oriented Linear Graphs*, pages 149–163. Birkhäuser Boston, Boston, MA, 1987. doi:10.1007/978-0-8176-4842-8_12.

35 Luca Q. Zamboni. On Christoffel words and their lexicographic array. *CoRR*, abs/2409.07974, 2024. doi:10.48550/arXiv.2409.07974.