Totally Unclustered BWT Images of Any Length over Non-Binary Alphabets

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

Estéban Gabory
Institute of Computer Science, University of Wrocław, Poland

Giuseppe Romana
Dipartimento di Matematica e Informatica, Università di Palermo, Italy

Marinella Sciortino
Dipartimento di Matematica e Informatica, Università di Palermo, Italy

Abstract

We prove that for every integer $n > 0$ and for every alphabet Σ_k of size $k \geq 3$, there exist words of length n whose Burrows–Wheeler Transform (BWT) is totally unclustered, i.e., it consists of exactly n runs with no two consecutive equal symbols. These words represent the worst-case behavior of the clustering effect of the BWT. We also establish a lower bound on their number. This contrasts with the binary case, where the existence of infinitely many totally unclustered BWT images is still an open problem, related to Artin's conjecture on primitive roots.

2012 ACM Subject Classification Mathematics of computing → Combinatorics; Theory of computation → Data structures design and analysis

Keywords and phrases Burrows–Wheeler Transform, BWT-runs, Repetitiveness Measure, Clustering Effect, Generalized de Bruijn Words

Digital Object Identifier 10.4230/LIPIcs.CPM.2026.13

Funding *Gabriele Fici*: Supported by MIUR project PRIN 2022 APML – 20229BCXNW.
Estéban Gabory: Partially supported by MIUR project PRIN 2022 APML – 20229BCXNW. Partially supported by the Polish National Science Centre grant number 2023/51/B/ST6/01505.
Giuseppe Romana: Supported by the project ACoMPA – Algorithmic and Combinatorial Methods for Pangenome Analysis (CUP B73C24001050001) funded by the NextGeneration EU programme PNRR MUR M4 C2 Inv. 1.5 – Project ECS00000017 Tuscany Health Ecosystem (Spoke 6), CUP Master B63C22000680007, and by the INdAM – GNCS Project CUP_E53C25002010001.
Marinella Sciortino: Partially supported by the project ACoMPA – Algorithmic and Combinatorial Methods for Pangenome Analysis (CUP B73C24001050001) funded by the NextGeneration EU programme PNRR MUR M4 C2 Inv. 1.5 – Project ECS00000017 Tuscany Health Ecosystem (Spoke 6), CUP Master B63C22000680007, and by the INdAM – GNCS Project CUP_E53C25002010001.

1 Introduction

The Burrows–Wheeler Transform (BWT) is a transformation on words introduced in 1994 [4]. When applied to a text (also called a word in the context of this paper), it produces a permutation of the characters obtained by concatenating the last character of each cyclic rotation of the text, after sorting all such rotations in lexicographic order. Since two rotations of the same word yield the same BWT, this transform can be naturally viewed as a mapping from the set of necklaces over a finite ordered alphabet $\Sigma_k = \{0, 1, \ldots, k-1\}$ (i.e., equivalence classes of words under rotation) to the set of words of Σ_k^*. The BWT is a widely used tool in data compression and indexing, because it is very likely to produce long runs of identical

37th Annual Symposium on Combinatorial Pattern Matching (CPM 2026).
Editors: Philip Bille and Nicola Prezza; Article No. 13; pp. 13:1–13:17
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

consecutive symbols (*BWT-runs*) when the input text is highly repetitive. This effect is usually called the *clustering effect* of the BWT. From a theoretical perspective, counting the number of BWT-runs provides a natural measure of repetitiveness of the input text and a suitable parameter for evaluating the performance of compressed indexing structures for highly repetitive data. This measure is also sensitive even to simple combinatorial operations, which can drastically change the number of BWT-runs [14, 1, 11, 15, 12]. The relation between the number of BWT-runs and other measures of repetitiveness, in particular those based on dictionary compressors, has been investigated [27, 19]. Moreover, combinatorial properties of texts that are maximally compressible through the BWT, namely those having the minimum number of BWT-runs, have also been studied [25, 7, 20, 10]. From an application perspective, the remarkable clustering effect of the BWT has been recognized as the key to its compression power [28], and it is effectively exploited in both text compression and compressed indexing when combined with run-length encoding techniques [26, 8, 13].

However, it is interesting to note that, while the BWT of a text typically exhibits some degree of clustering, there exist texts for which no such clustering occurs. In this work, we focus precisely on this phenomenon, which so far has been the least explored: the existence of words whose BWT is *totally unclustered*, i.e., such that no two consecutive characters are equal. These totally unclustered BWT images correspond to a worst-case scenario for compression and indexing methods based on run-length BWT, which achieve space and time efficiency when the number of BWT runs is much smaller than the text length.

This problem was previously studied by Mantaci et al. [24] in the case of binary alphabets, revealing a connection with the generators of the multiplicative group $\mathbb{Z}_n^\times$ of integers modulo n, as stated in the following result:

▶ **Theorem 1** ([24]). *There exists a word of length $2m$ with a totally unclustered BWT if and only if $2m+1$ is an odd prime and 2 generates the cyclic group $\mathbb{Z}_{2m+1}^\times$.*

However, the question of whether infinitely many binary words with totally unclustered BWT exist is left open in [24]. This problem is tightly connected to the still-open Artin's conjecture on the existence of primitive roots modulo infinitely many prime numbers.

Here, we move beyond the binary setting and prove that the existence of totally unclustered BWT images is guaranteed for every length as soon as the alphabet contains at least three letters. In particular, we prove the following:

▶ **Theorem 2.** *For every $k \geq 3$, and for every $n \geq k$, there exists a word u of length n over Σ_k such that every letter of Σ_k appears at least once and $BWT(u)$ is totally unclustered.*

Our proof is constructive, as we explicitly show how to build words of any length whose BWT is totally unclustered. More in detail, our approach builds upon the structure of generalized de Bruijn words introduced in [9]. The key point is that generalized de Bruijn words have a BWT consisting of consecutive blocks, each of which is a permutation of the alphabet. As proved in [9], generalized de Bruijn words are precisely the Hamiltonian cycles of the generalized de Bruijn graphs, introduced in the early 80s independently by Imase and Itoh [18], and by Reddy, Pradhan, and Kuhl [30] (see also [6]) in the context of network design.

In particular, we prove the following result:

▶ **Theorem 3.** *For every integer $m \geq 1$, and $k \geq 3$, there exists a generalized de Bruijn word of length $n = km$ over Σ_k with totally unclustered BWT.*

Here, we provide a new algorithmic strategy that exploits combinatorial properties of generalized de Bruijn graphs, making it possible to remove equal letters between consecutive blocks of the BWT, thereby obtaining words of length $n = km$ whose BWT images are totally unclustered. We then show how to extend this procedure to every length, not just multiples of the size of the alphabet. To the best of our knowledge, this is the first work exhibiting words over fixed-size alphabets with $n - o(n)$ BWT-runs for all n.

A de Bruijn word of order d over Σ_k is a word of length k^d that contains every string of length d exactly once as a circular factor. Since generalized de Bruijn words reduce to ordinary ones whenever n is a power of k, Theorem 3 guarantees the existence of ordinary de Bruijn words with totally unclustered BWTs for such n. Moreover, we prove that for such words, the ratio between the number of runs in the BWT and the number of runs in the lexicographically smallest rotation is exactly $\frac{k}{k-1}$, when $k \geq 3$. It has been proved in [24, Theorem 3.3] that this ratio, also called *clustering ratio*, is always at most 2. In this sense, our result shows that the BWT always increases the number of runs of the lexicographically smallest rotation of the de Bruijn word to which it is applied. This was previously known only for binary alphabets, although it has been proven that the clustering ratio is never equal to 2 for binary de Bruijn words, when $d \geq 3$ [24, Theorem 5.2].

Beyond the existence of words with totally unclustered BWT, we also investigate counting aspects: we provide a lower bound of $\Omega(2^{\frac{n}{3}}/n)$ on the number of words of length n, over any alphabet of size at least 3, with totally unclustered BWT. We conclude by examining a special case connected to Artin's conjecture, showing that certain highly structured words of length $n = km$ are BWT images precisely when $km + 1$ is prime and k is a primitive root modulo $km + 1$.

2 Preliminaries

We begin by introducing some preliminary definitions. For a thorough introduction, we refer the reader to [5] and [22].

Words and Necklaces

Let $\Sigma_k = \{\mathtt{0}, \mathtt{1}, \ldots, k-1\}$, with $k > 1$, denote the integer *alphabet* of size k, whose elements are called *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 $(|w|_{\mathtt{0}}, \ldots, |w|_{k-1})$ is the *Parikh vector* of w. For a word w, the *n-th power* of w is the word w^n obtained by concatenating n copies of w. A word w is *primitive* if, for any non-empty word v and integer n, $w = v^n$ implies $v = w$ and $n = 1$. For any integer $k > 1$, by *alphabet permutation* we mean any string w of length $|\Sigma_k|$ satisfying $|w|_i = 1$ for all $i \in \{\mathtt{0}, \mathtt{1}, \ldots, k-1\}$. In other words, it is any word whose Parikh vector is $(1, 1, \ldots, 1)$. We call a concatenation of one or more alphabet permutations over Σ_k an *alphabet-permutation power*.

▶ **Example 4.** The word $w = \mathtt{210201102102120}$ is an alphabet-permutation power over Σ_3.

Let $w = w_0 \cdots w_{n-2}w_{n-1}$ be a word of length $n > 0$. Two words w and w' are *conjugates* if $w = uv$ and $w' = vu$ for some words u and v, i.e., if w' is a cyclic rotation of w. A *necklace* (resp. *aperiodic necklace*) $[w]$ is the conjugacy class of words (resp. of primitive words). Observe that a necklace $[w]$ is aperiodic if and only if $|[w]| = |w|$, i.e., all the conjugates of w are distinct.

▶ **Example 5.** Consider the words $u = \texttt{1100}$ and $v = \texttt{1010}$. The necklace $[u] = [\texttt{1100}] = \{\texttt{1100},\texttt{0110},\texttt{0011},\texttt{1001}\}$ is aperiodic, while $[v] = [\texttt{1010}] = \{\texttt{1010},\texttt{0101}\}$ is not.

Burrows–Wheeler Transform and Standard Permutation

The *Burrows–Wheeler table* (BWT table) of a word u is the square table whose rows are the cyclic rotations of u, taken in ascending lexicographic order. Let us denote by F and L, respectively, the first and the last column of the BWT table of u, read from top to bottom. We have that $F = \texttt{0}^{n_0}\texttt{1}^{n_1}\cdots(k-1)^{n_{k-1}}$, where $(n_0, \ldots, n_{k-1})$ is the Parikh vector of u; L, instead, is the *Burrows–Wheeler Transform* (BWT) of u, denoted $BWT(u)$. From its definition, it follows that for two words $u, v \in \Sigma^n$ it holds that $BWT(u) = BWT(v)$ if and only if u and v are conjugates. Since the action of the BWT on a necklace $[u]$ is the same as its action on any conjugate of u, we use the notation $BWT(u)$ or $BWT([u])$ indifferently. The BWT is therefore an injective map from the set of necklaces to the set of words. A word is a *BWT image* if it is the BWT of some necklace.

F																	L
0	0	1	1	0	2	1	0	0	1	1	2	0	2	1	2	2	2
0	0	1	1	2	0	2	1	2	2	2	0	0	1	1	0	2	1
0	1	1	0	2	1	0	0	1	1	2	0	2	1	2	2	2	0
0	1	1	2	0	2	1	2	2	2	0	0	1	1	0	2	1	0
0	2	1	0	0	1	1	2	0	2	1	2	2	2	0	0	1	1
0	2	1	2	2	2	0	0	1	1	0	2	1	0	0	1	1	2
1	0	0	1	1	2	0	2	1	2	2	2	0	0	1	1	0	2
1	0	2	1	0	0	1	1	2	0	2	1	2	2	2	0	0	1
1	1	0	2	1	0	0	1	1	2	0	2	1	2	2	2	0	0
1	1	2	0	2	1	2	2	2	0	0	1	1	0	2	1	0	0
1	2	0	2	1	2	2	2	0	0	1	1	0	2	1	0	0	1
1	2	2	2	0	0	1	1	0	2	1	0	0	1	1	2	0	2
2	0	0	1	1	0	2	1	0	0	1	1	2	0	2	1	2	2
2	0	2	1	2	2	2	0	0	1	1	0	2	1	0	0	1	1
2	1	0	0	1	1	2	0	2	1	2	2	2	0	0	1	1	0
2	1	2	2	2	0	0	1	1	0	2	1	0	0	1	1	2	0
2	2	0	0	1	1	0	2	1	0	0	1	1	2	0	2	1	2
2	2	2	0	0	1	1	0	2	1	0	0	1	1	2	0	2	1

Figure 1 BWT table of the necklace [011021001120212220]. The BWT of such a necklace can be read from top to bottom in the last column L, highlighted in yellow.

▶ **Example 6.** Let us consider the necklace $[u]$, where $u = \texttt{011021001120212220}$ of length 18. The BWT table of u is depicted in Fig. 1, from which it follows that $BWT(u) = \texttt{210012210012210021}$.

The *run-length encoding* of a word $w \in \Sigma^n$, denoted $\mathrm{RLE}(w)$, is the sequence $(w_i, \ell_i)_{\{1 \le i \le d\}} \in (\Sigma \times \mathbb{N})^d$, where $w = w_1^{\ell_1} \cdots w_d^{\ell_d}$, and $w_i \neq w_{i+1}$ for every $1 \le i < d$. If $w = BWT(u)$ for a necklace $[u]$, we write $r(u) = |\mathrm{RLE}(w)|$. The *clustering ratio* of a necklace $[u]$ is the ratio between the number of equal-letter runs in the BWT of $[u]$ and its lexicographically smallest rotation v in $[u]$, i.e., $\frac{r(u)}{|\mathrm{RLE}(v)|}$[1]. It has been proved that for any word, the clustering ratio is at most 2 [24, Theorem 3.3]. A word w is *totally unclustered* if $|\mathrm{RLE}(w)| = |w|$. We say that a word u (resp. a necklace $[u]$) has a *totally unclustered BWT* if $BWT(u)$ is totally unclustered.

[1] Note that $|\mathrm{RLE}(v)| \le |\mathrm{RLE}(u)| \le |\mathrm{RLE}(v)| + 1$. This follows from the fact that any rotation u can split at most one circular run of the necklace into two linear runs.

▶ **Example 7.** Consider the necklace $[u] = [\texttt{00012110111222}]$. Its Burrows–Wheeler Transform is the word $w = BWT(u) = \texttt{20101201012121}$. Observe that w is totally unclustered. Moreover, since u is the smallest word in lexicographical order in $[u]$, the clustering ratio is $\frac{r(u)}{|\,\mathrm{RLE}(u)|} = \frac{14}{7} = 2$.

A *permutation* σ on $\{0, \ldots, n-1\}$ is determined by the images of its elements and can be represented in *two-line notation* as $\begin{pmatrix} 0 & \ldots & n-1 \\ \sigma(0) & \ldots & \sigma(n-1) \end{pmatrix}$. It can also be defined in terms of its *cycle decomposition* as a sequence of tuples $c_0c_1 \cdots c_\ell$ partitioning $\{0, \ldots, n-1\}$ where, for all $j \in \{0, \ldots, \ell\}$, $c_j = (i_j, \sigma(i_j), \ldots, \sigma^{t_j}(i_j))$, with $i_j \in \{0, \ldots, n-1\}$ and $t_j \geq 0$ the smallest integer for which $\sigma^{t_j+1}(i_j) = i_j$. Such a decomposition is unique up to reordering of the cycles [31]. The *standard permutation* of a word $w = w_0w_1 \cdots w_{n-1}$, $w_i \in \Sigma_k$, is the permutation $\pi_w : \{0, 1, \ldots, n-1\} \to \{0, 1, \ldots, n-1\}$ such that $\pi_w(i) < \pi_w(j)$ if and only if $w_i < w_j$ or $w_i = w_j$ and $i < j$. In other words, in two-line notation, π_w orders distinct letters of w lexicographically, and equal letters by occurrence order, starting from 0. In the context of BWT (and its extensions to a multiset of words [23]), the standard permutation of the output word is also called the LF-mapping. The *inverse standard permutation* π_w^{-1} of a word w, written in two-line notation, can be obtained by listing in left-to-right order the positions of 0 in w, then the positions of 1, and so on. The inverse standard permutation of a word produced as output by the BWT (or by its extensions to a multiset of words) is also called FL-mapping.

The following lemma is a direct consequence of the definition of BWT and [23, Theorem 13].

▶ **Lemma 8.** *An n-length word w is a BWT of some aperiodic necklace $[u]$ if and only if π_w (or equivalently, π_w^{-1}) is a cycle of length n.*

The following examples illustrate the two cases in which the standard permutation is not a single cycle: words that are not BWT images of any necklace, and words that are BWT images of periodic necklaces.

▶ **Example 9.** Let us consider the word $w = \texttt{210201102102120}$ used in Example 4. The standard permutation and its cycle decomposition are as follows:

$$\begin{aligned}\pi_w &= \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\ 10 & 5 & 0 & 11 & 1 & 6 & 7 & 2 & 12 & 8 & 3 & 13 & 9 & 14 & 4 \end{pmatrix} \\ &= (0, 10, 3, 11, 13, 14, 4, 1, 5, 6, 7, 2)(8, 12, 9).\end{aligned}$$

The inverse standard permutation and its cycle decomposition of w are as follows:

$$\begin{aligned}\pi_w^{-1} &= \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\ 2 & 4 & 7 & 10 & 14 & 1 & 5 & 6 & 9 & 12 & 0 & 3 & 8 & 11 & 13 \end{pmatrix} \\ &= (0, 2, 7, 6, 5, 1, 4, 14, 13, 11, 3, 10)(8, 9, 12).\end{aligned}$$

By using Lemma 8, w is not the image of any aperiodic necklace under the BWT.

▶ **Example 10.** Let us consider the periodic necklace $[u]$, where $u = \texttt{012101210121} = (\texttt{0121})^3$. The reader can verify that $BWT(\texttt{0121}) = \texttt{1201}$ and $w = BWT(u) = \texttt{111222000111}$. The standard permutation and its cycle decomposition are as follows:

$$\begin{aligned}\pi_w &= \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 3 & 4 & 5 & 9 & 10 & 11 & 0 & 1 & 2 & 6 & 7 & 8 \end{pmatrix} \\ &= (0, 3, 9, 6)(1, 4, 10, 7)(2, 5, 11, 8).\end{aligned}$$

Generalized De Bruijn Graphs and Words

Let $G = (V, E)$ be a finite directed *graph*, where V is the set of *vertices* and $E \subseteq V \times V$ is the set of *edges*. Each edge $e = (v_1, v_2) \in E$ is directed from its *source vertex* $s(e) = v_1$ to its target vertex $t(e) = v_2$. We write $\mathrm{indeg}(v) = |\{e \in E \mid \ t(e) = v\}|$, and $\mathrm{outdeg}(v) = |\{e \in E \mid \ s(e) = v\}|$ for every $v \in V$. A *path* γ in the graph G is any sequence of edges $e_1, e_2, \ldots, e_k$ verifying that $t(e_i) = s(e_{i+1})$ for all $i \in \{1, \ldots, k-1\}$; the path γ is further called *cycle* if $s(e_1) = t(e_k)$.

Given a finite directed graph $G = (V, E)$, the directed *line graph* (or *edge graph*) $\mathcal{L}G = (E, E')$ of G is a graph having as vertices the edges of G, and as edges the set

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

A directed graph is *Hamiltonian* if it has a *Hamiltonian cycle*, i.e., one that traverses each vertex 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 $\mathrm{indeg}(v) = \mathrm{outdeg}(v)$ for all vertices v. If a graph is Eulerian, its line graph is Hamiltonian.

For all integers $k, d \geq 1$, the *de Bruijn graph* $\mathrm{DB}(k, d)$ is the directed graph where each vertex v is a string of length d over Σ_k and there exists an edge $e = (iw, wj)$ for each $i, j \in \Sigma_k$ and $w \in \Sigma_k^{d-1}$. As is well known, de Bruijn graphs are both Eulerian and Hamiltonian, and one has $\mathcal{L}\,\mathrm{DB}(k, d) = \mathrm{DB}(k, d+1)$.

The *generalized de Bruijn graph* $\mathrm{GDB}(k, m)$ is a graph having as vertices $\{0, 1, \ldots, m-1\}$, and for every vertex v there is an edge from v to $kv + i \mod m$ for every $i = 0, 1, \ldots, k-1$. Observe that for every $d > 0$, it holds[2] that $\mathrm{GDB}(k, k^d) = \mathrm{DB}(k, d)$. Generalized de Bruijn graphs have been introduced independently by Imase and Itoh [18], and by Reddy, Pradhan, and Kuhl [30] (see also [6]). They are Eulerian for every m and k. The line graph of $\mathrm{GDB}(k, m)$ is $\mathrm{GDB}(k, km)$ [21]. More precisely, one has the following:

▶ **Lemma 11** ([21]). *Given k, m integers, the graph $\mathrm{GDB}(k, km)$ is isomorphic to the line graph of $\mathrm{GDB}(k, m)$, where the vertex corresponding to an edge $(i, ki+j \mod m)$ is associated to the edge with label $ki + j \mod km$.*

The Eulerian cycles of $\mathrm{GDB}(k, m)$ correspond to the Hamiltonian cycles of $\mathrm{GDB}(k, km)$.

A *de Bruijn word* of order d on Σ_k is a necklace $[u]$ of length k^d such that every string in Σ_k^d occurs exactly once as a circular 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 table of a de Bruijn word, the Burrows–Wheeler transform of a de Bruijn word is an alphabet-permutation power (cf. [16]).

De Bruijn words were generalized in several different ways (see, e.g., [3, 2, 29]). We use here the definition from [9]: a necklace $[u]$ of length $n = km$ over the alphabet Σ_k is a *generalized de Bruijn word* if $BWT(u)$ is an alphabet-permutation power. Generalized de Bruijn words are aperiodic necklaces.

▶ **Example 12.** Consider the word $u =$ `011021001120212220` from Example 6. The necklace $[u]$ is a generalized de Bruijn word. Indeed, $BWT(u) =$ `210012210012210021` is an alphabet-permutation power.

Observe that not all alphabet-permutation powers are BWT images, as witnessed by the word in Example 9.

[2] Up to renaming integer labels with their base-k representations in d digits.

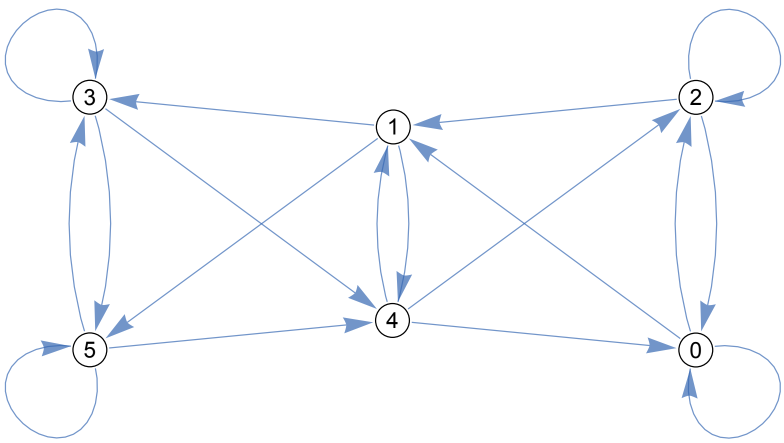

Figure 2 Generalized de Bruijn graph GDB$(3,6)$. The graph has six vertices, labeled $0,1,2,3,4,5$. From every vertex m, there are directed edges to vertices $(3m+i) \bmod 6$ for each $i \in \{0,1,2\}$.

The following theorem establishes a direct correspondence between generalized de Bruijn words and Hamiltonian cycles in generalized de Bruijn graphs.

▶ **Theorem 13** ([9]). *A necklace $[u]$ of length $n = km$ over Σ_k is a generalized de Bruijn word if and only if $\pi^{-1}_{BWT(u)}$ is a Hamiltonian cycle of* GDB(k, km)*. When $n = k^d$ for some integer $d \geq 1$, generalized de Bruijn words of length n over Σ_k coincide with ordinary de Bruijn words of order d on Σ_k.*

The next example shows a generalized de Bruijn word whose BWT is a totally unclustered alphabet-permutation power, also providing an instance of the statement of Theorem 13.

▶ **Example 14.** The necklace $[u]$, where $u =$ `220011021002211201` is a generalized de Bruijn word of length $3 \cdot 6 = 18$, since $BWT(u) =$ `210` $\cdot$ `210` $\cdot$ `210` $\cdot$ `210` $\cdot$ `210` $\cdot$ `210` is an alphabet-permutation power. Moreover, $\rho(u) = 18 = |BWT(u)|$, so $[u]$ has a totally unclustered BWT. Note that the standard permutation $\pi_{BWT(u)}$ is a cycle of length 18:

$$\pi_{BWT(u)} = (0, 12, 16, 11, 3, 13, 10, 9, 15, 17, 5, 1, 6, 14, 4, 7, 8, 2).$$

The inverse permutation is:

$$\pi^{-1}_{BWT(u)} = (0, 2, 8, 7, 4, 14, 6, 1, 5, 17, 15, 9, 10, 13, 3, 11, 16, 12).$$

3 Main Result

In this section, we establish the existence of necklaces (and, hence, of classes of words) of any length with totally unclustered BWT over alphabets of size at least three. Recall from [25, Proposition 2] that if a necklace has totally unclustered BWT, then it must be aperiodic. Our approach is based on generalized de Bruijn words and their representation as Hamiltonian cycles in generalized de Bruijn graphs. We first show how to construct such necklaces of length that is a multiple of the alphabet size, and then how to extend the construction to arbitrary lengths.

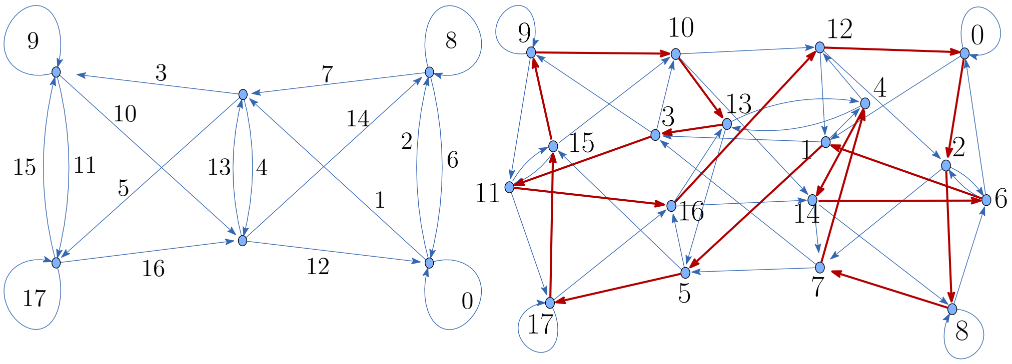

Figure 3 The graph GDB(3, 6) (left) and the graph $\mathcal{L}\,$GDB(3, 6) = GDB(3, 18) (right). The ith edge in GDB(3, 6) corresponds to the ith vertex in GDB(3, 18). The Hamiltonian cycle corresponding to $\pi^{-1}_{BWT(u)} = (0, 2, 8, 7, 4, 14, 6, 1, 5, 17, 15, 9, 10, 13, 3, 11, 16, 12)$ (Example 14) is highlighted in red on GDB(3, 18). It corresponds to an Eulerian cycle on GDB(3, 6). Notice that the edges of the form $(i, 3i + j \mod 6)$ in GDB(3, 6) (where the vertices are labeled as in Figure 2) are labeled $3i + j \mod 18$ (Lemma 11).

3.1 Constructing totally unclustered BWT images from generalized de Bruijn words

We begin by proving that for every integer $m > 0$, and $k \geq 3$, there exists a generalized de Bruijn word of length $n = km$ over Σ_k with totally unclustered BWT. The key insight is that adjacent equal letters between consecutive alphabet-permutation blocks can be systematically eliminated by rerouting Hamiltonian cycles in the corresponding de Bruijn graph. This yields generalized de Bruijn words with totally unclustered BWT.

Let $[u]$ be a generalized de Bruijn word, and let us write $w = BWT(u)$. By definition, w is an alphabet-permutation power, hence a sequence of length-k blocks that are permutations of Σ_k. Clearly, a repeated letter can occur only at the boundary between two consecutive such blocks. In other words, u has a totally unclustered BWT if and only if $w_{ki+(k-1)} \neq w_{k(i+1)}$ for every $0 \leq i \leq m - 2$. On the contrary, if $w_{ki+(k-1)} = w_{k(i+1)}$, we say that w has a *tie* at block i. Ties for w can also be characterized by the inverse standard permutation of w, as shown in the following lemma.

▶ **Lemma 15.** *Let $[u]$ be a generalized de Bruijn word of length $n = km$ over Σ_k, $w = BWT(u)$, and $\sigma = \pi_w^{-1}$. The word w has a tie at block $i \in [0 \ldots m-2]$ if and only if for some $j \in \{0, \ldots, k-1\}$, one has:*

$$\sigma(jm + i) = ki + k - 1, \text{ and}$$
$$\sigma(jm + i + 1) = k(i + 1).$$

Proof. Let $w = BWT(u)$. Consider the permutation $\sigma = \pi_w^{-1}$. It can be understood as follows: writing the multiset of letters of w in lexicographic order, where equal letters are sorted by occurrence order in w, the position in w of the i-th letter is $\sigma(i)$. In particular, since the Parikh vector of w is $(m, \ldots, m)$, the positions of 0 in w are $\sigma(0), \ldots, \sigma(m-1)$, the positions of 1 are $\sigma(m), \ldots, \sigma(2m-1)$, and so on, until the last letter, which occurs at positions $\sigma((k-1)m), \ldots, \sigma(km-1)$. In particular, there is a tie in w at block i if and only if, for some $j \in \{0, \ldots, k-1\}$ the jth letter of Σ_k occurs at positions $ki + k - 1$ and $k(i + 1)$, and in that case they are respectively the $(i + 1)$-th and the $i + 2$-th occurrences of this letter. In other words, we have $\sigma(jm + i) = ki + k - 1$ and $\sigma(jm + i + 1) = k(i + 1)$, as claimed. ◀

The following lemma provides conditions under which Hamiltonian cycles can be rerouted to avoid a specified edge. We will use it to break ties in BWT images of generalized de Bruijn words by rerouting the corresponding Hamiltonian cycles so that they avoid one of the edges responsible for the tie.

▶ **Lemma 16.** *Let G be a directed Hamiltonian graph, and $A, B \subseteq V(G)$ such that:*

1. *both A and B have cardinality at least 3;*
2. *every pair $(x, y) \in A \times B$ is an edge of G;*
3. *every edge starting at a vertex in A ends at a vertex in B.*

Then, if γ is a Hamiltonian cycle for G visiting an edge $e \in A \times B$, it can be rerouted into a Hamiltonian cycle γ' such that: (i) γ' does not visit e, and (ii) if $e' \in E(G) \setminus (A \times B)$, γ' visits e' if and only if γ visits e'.

Proof. Let G be a directed Hamiltonian graph, A and B two subsets of $V(G)$ satisfying the conditions of the statement, and γ a Hamiltonian cycle visiting $e = (x, y) \in A \times B$. Since γ is a cycle, we can write it as a sequence of vertices starting with y and ending with x. Sets A and B have cardinality at least 3, and every vertex from A must be followed by a vertex from B. Hence, one can write $\gamma = \gamma_1 \cdot \gamma_2 \cdot \gamma_3$ where each γ_i starts with a vertex from B and ends with a vertex from A, for every $i \in \{1, 2, 3\}$. Since each $(x, y) \in A \times B$ is an edge, the sequence $\gamma' = \gamma_1 \cdot \gamma_3 \cdot \gamma_2$ is still a Hamiltonian cycle for G, which does not visit (x, y). Furthermore, γ and γ' differ only on edges in $A \times B$, and the statement follows. ◀

▶ Remark 17. Note that Lemma 16 does not require having $A \cap B = \emptyset$.

▶ Remark 18. If G is the line graph of an Eulerian graph $\tilde{G}$, Lemma 16 has a simpler interpretation: the set A (resp. B) corresponds to the set of all incoming (resp. all outgoing) edges for a fixed vertex v having indegree and outdegree at least 3, and the edge e to be avoided corresponds to a sequence of two edges (x, v) and (v, y). In that case, the cycle γ corresponds to an Eulerian cycle on $\tilde{G}$, which by definition must visit v at least 3 times, and can be written as a concatenation of 3 cycles starting and ending at v. In that case, the rerouting performed corresponds to a permutation of the order in which those cycles are traversed. We presented the Hamiltonian version of the result because the correspondence of Theorem 13 is more direct from the Hamiltonian viewpoint; however, all related statements can be reformulated in terms of Eulerian graphs.

The following lemma shows that the rightmost tie in the BWT of a generalized de Bruijn word can always be removed.

▶ **Lemma 19.** *Let $[u]$ be a generalized de Bruijn word of length $n = km$ over Σ_k, with $k \geq 3$, and $w = BWT(u)$. Assume that i_0 is the rightmost tie for w. Then, there exists w' such that: (i) $w' = BWT(u')$ for some generalized de Bruijn word $[u']$ of length $n = km$ over Σ_k; (ii) $w'_j = w_j$ for every $j \notin \{ki_0, ki_0 + 1, \ldots, ki_0 + k - 1\}$; (iii) w' has no tie at block i for any $i \geq i_0$.*

Proof. Suppose i_0 is the rightmost tie of w. From the hypothesis, it follows that $w_{ki_0+k-1} = w_{k(i_0+1)}$, and from Lemma 15, there is a fixed $j_0 \in \{0, \ldots, k-1\}$ such that $\sigma(j_0 m + i_0) = ki_0 + k - 1$ and $\sigma(j_0 m + i_0 + 1) = k(i_0 + 1)$, where j_0 depends on the letter $w_{ki_0+k-1} = w_{k(i_0+1)}$.

By Lemma 8, the permutation $\sigma = \pi_w^{-1}$ is a cycle, and by Theorem 13, this cycle corresponds to the consecutive vertex labels of a Hamiltonian cycle of $G = \mathrm{GDB}(k, km)$. Consider the sets $A_{i_0} = \{i_0, m + i_0, \ldots, (k-1)m + i_0\} \subseteq V(G)$, $B_{i_0} = \{ki_0, ki_0 + 1, \ldots, ki_0 + k - 1\} \subseteq V(G)$. Note that the edge $(j_0 m + i_0, ki_0 + k - 1)$ is visited by the cycle corresponding to σ and that it is an edge from A_{i_0} to B_{i_0}. Those sets both have cardinality $k \geq 3$, and by construction of G, they satisfy the conditions of Lemma 16. We can then obtain a Hamiltonian cycle of G, corresponding to a permutation σ', with $\sigma'(j_0 m + i_0) \neq k i_0 + k - 1$, hence w' has no tie at i_0, and $\sigma'(\ell) = \sigma(\ell)$ for every $\ell \notin A_{i_0}$ (importantly, we still have $\sigma(j_0 m + i_0 + 1) = k(i_0 + 1)$, so the conditions of Lemma 15 cannot hold for some $j \neq j_0$), and the words w and w' differ only at block i_0 (namely, at positions $k i_0, \ldots, k i_0 + k - 1$). Since i_0 was the rightmost tie in w, we deduce that w' has no tie at block $i \geq i_0$. ◀

We can now prove our main result:

▶ **Theorem 3.** *For every integer $m \geq 1$, and $k \geq 3$, there exists a generalized de Bruijn word of length $n = km$ over Σ_k with totally unclustered BWT.*

Proof. By Theorem 13, generalized de Bruijn words are in bijection with Hamiltonian cycles in $\mathrm{GDB}(k, km)$, or equivalently, with Eulerian cycles in $\mathrm{GDB}(k, m)$. Since $\mathrm{GDB}(k, m)$ is Eulerian for every k, m, there is at least one generalized de Bruijn word u of length km over Σ_k. If $BWT(u)$ has no tie, then u is a generalized de Bruijn with totally unclustered BWT. Otherwise, let i_0 be the rightmost tie for w. By Lemma 19, there exists $w' = BWT([u'])$, for some generalized de Bruijn word $[u]$, such that either w' has no tie, and therefore $[u']$ is totally unclustered, or its rightmost tie is at block $i_0' < i_0$, and recursively apply at most $i_0 + 1$ times the same approach until we eventually obtain a word with no ties, which is then the BWT of some generalized de Bruijn word that has a totally unclustered BWT. ◀

▶ **Example 20.** Let us consider the necklace $[u]$, where $u = \texttt{001022112}$. One has $w = BWT(u) = \texttt{201021120}$, hence $[u]$ is a generalized de Bruijn word of length km over Σ_k with $k = m = 3$. Since w has a tie at block $i = 1$, u does not have a totally unclustered BWT. One has $\sigma = \pi_w^{-1} = (0, 1, 3, 2, 8, 7, 4, 5, 6)$, and $\sigma(4) = \sigma(m + i) = 5 = k \cdot i + 2$, and $\sigma(5) = \sigma(m + i + 1) = 6 = k \cdot (i + 1)$. We have $A_1 = \{i, m + i, 2m + i\} = \{1, 4, 7\}$, $B_1 = \{3i, 3i + 1, 3i + 2\} = \{3, 4, 5\}$, and we write $\gamma = 560132874 = \gamma_1 \cdot \gamma_2 \cdot \gamma_3$, with $\gamma_1 = 5601$, $\gamma_2 = 3287$, $\gamma_3 = 4$. Then, $\gamma' = \gamma_1 \cdot \gamma_3 \cdot \gamma_2 = 560143287$ also corresponds to a cycle on $\mathrm{GDB}(3, 9)$, namely $\sigma' = (0, 1, 4, 3, 2, 8, 7, 5, 6)$, and $\sigma' = \pi_{w'}^{-1}$, with $w' = \texttt{201102120} = BWT(u')$, $u' = \texttt{001102212}$. The word w' still has one tie – observe, however, that the tie is now at block $i' = 0$, and that there are no ties at later blocks.

One has $\sigma'(3) = \sigma'(m + 0) = 2 = k \cdot i + 2$, and $\sigma'(4) = \sigma(m + 1) = 3 = k \cdot (i + 1)$. We have $A_0 = \{0, 3, 6\}$, $B_0 = \{0, 1, 2\}$, and we write $\gamma^{(2)} = 287560143 = \gamma_1^{(2)} \cdot \gamma_2^{(2)} \cdot \gamma_3^{(2)}$ with $\gamma_1^{(2)} = 28756$, $\gamma_2^{(2)} = 0$ $\gamma_3^{(2)} = 143$. Then, $\gamma'^{(2)} = \gamma_1^{(2)} \cdot \gamma_3^{(2)} \cdot \gamma_2^{(2)} = 287561430$ corresponds to a cycle of $\mathrm{GDB}(3, 9)$, namely $\sigma'^{(2)} = (0, 2, 8, 7, 5, 6, 1, 4, 3)$, and $\sigma'^{(2)} = \pi_{w'^{(2)}}^{-1}$, with $w'^{(2)} = \texttt{120102120} = BWT(u'^{(2)})$, where $u'^{(2)} = [\texttt{002212011}]$. The necklace $[u'^{(2)}]$ is a de Bruijn word with totally unclustered BWT and its clustering ratio is $\frac{3}{2}$.

As a direct consequence of Theorem 3, we obtain the following result for ordinary de Bruijn words.

▶ **Corollary 21.** *For every $k \geq 3$, and every $d \geq 1$, there exists a de Bruijn word $[u]$ over Σ_k of order d (and hence length k^d) with totally unclustered BWT.*

Corollary 21 shows a clear separation from the binary case, where de Bruijn words with totally unclustered BWT do not exist [24, Theorem 5.2].

Moreover, in the case of de Bruijn words with totally unclustered BWT, the clustering ratio depends only on k, as shown in the following proposition.

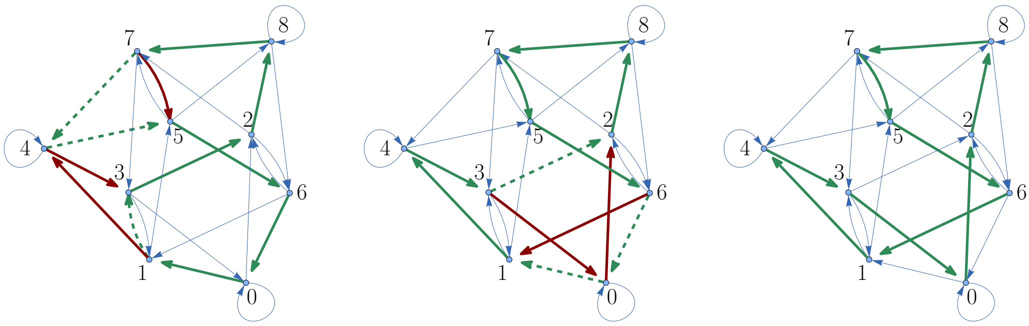

Figure 4 Illustration of the process on the de Bruijn graph GDB$(3,9)$, as described in Example 20. In the left subfigure, the green cycle corresponds to the initial permutation σ, and the dashed green edges (resp. red edges) indicate the edges to remove (resp. to add) to obtain the updated permutation. In the middle subfigure, the green cycle corresponds to $\sigma^{(2)}$, and the dashed green edges (resp. red edges) indicate the edges to remove (resp. to add) to obtain $\sigma'^{(2)}$. The right subfigure shows the final cycle $\sigma'^{(2)}$ in green, which corresponds to the inverse standard permutation of a generalized de Bruijn word with a totally unclustered BWT.

▶ **Proposition 22.** *Let $[u]$ be a de Bruijn word over Σ_k of order d with totally unclustered BWT, for some $k \geq 3$ and $d \geq 1$. If v is the lexicographically smallest rotation in $[u]$, then $|\operatorname{RLE}(v)| = k^{d-1}(k-1)$, and $r(v)/|\operatorname{RLE}(v)| = k/(k-1)$.*

Proof. Recall that $[u]$ is a de Bruijn word of length k^d. Clearly, $|\operatorname{RLE}(v)| = 1 + |\{0 \leq i \leq k^d - 2,\ v[i] \neq v[i+1]\}|$. Since $[u]$ is a de Bruijn word, each string from Σ_k^2 occurs exactly k^{d-2} times as a circular factor of $[u]$, and Σ_k^2 contains $k^2 - k$ strings having two distinct letters. Finally, since v is minimal among rotations of $[u]$, its first and last letters must be distinct, hence exactly one of the circular occurrences of a pair of distinct letters is not an occurrence in v. Wrapping up, we obtain $|\operatorname{RLE}(v)| = k^{d-2}(k^2 - k) = k^{d-1}(k-1)$. Now, since v has totally unclustered BWT, which means that $r(v) = |v| = k^d$, we conclude that $r(v)/|\operatorname{RLE}(v)| = k/(k-1)$. ◀

▶ Remark 23. Words that have a totally unclustered BWT are examples of words for which the BWT cannot be compressed by run-length encoding. Additionally, not only can a word have a totally unclustered BWT, but also the clustering ratio can be maximized, as witnessed by the word $u = 000211011222$, whose BWT is 201012012021, thus $[u]$ is a generalized de Bruijn word with totally unclustered BWT and the clustering ratio is 2. However, we do not know if this can happen for infinitely many lengths.

In the case of ordinary de Bruijn words over alphabets of size at least 3, Proposition 22 tells us that the clustering ratio cannot achieve the maximum value 2. However, it implies the existence of an infinite family of ternary words of increasing length whose BWT are totally unclustered and for which the clustering ratio equals 3/2, thus highlighting the unclustering effect of the BWT on this family of ternary words.

3.2 Extending to all lengths

After establishing the existence of totally unclustered BWT images for generalized de Bruijn words of length $n = km$, we extend the construction to arbitrary word lengths in the ternary case. In fact, we show how to insert or delete a single letter while preserving the cyclic structure of the standard permutation.

We start by introducing some notation. For $m > 0$, let P_m and Q_m be two sets defined as follows: if u is a ternary generalized de Bruijn word of length $n = 3m$, then $w = BWT(u) \in P_m$ (resp. $w \in Q_m$) if and only if $w_{3m-2} = 2$ (resp. if $w_{3m-3} = 2$).

▶ **Lemma 24.** *Let $m > 0$. The sets P_m and Q_m are non-empty and in bijection.*

Proof. Let u be a ternary generalized de Bruijn word of length $n = 3m$, and $w = BWT(u)$. Let $\sigma = \pi_w^{-1}$ be its inverse standard permutation, corresponding to a Hamiltonian cycle γ of GDB$(3, 3m)$. Let $A_m = \{m-1, 2m-1, 3m-1\}$ and $B_m = \{3m-3, 3m-2, 3m-1\}$, and write e_1 for the edge $(3m-1, 3m-2)$ and $e_2 = (3m-1, 3m-3)$. Notice also that the edge $(3m-1, 3m-1)$ is visited by no Hamiltonian cycle, since it is a self-loop. Now, γ must visit one edge among e_1 and e_2: in particular, it visits e_1 if and only if $w \in P_m$, and e_2 if and only if $w \in Q_m$. In both cases, since A_m and B_m satisfy the conditions of Lemma 16, we can obtain a Hamiltonian cycle γ' visiting e_2 (resp. e_1) if and only if γ visits e_1 (resp. e_2), hence a word w' which is the BWT image of a de Bruijn word and such that $w' \in Q_m$ (resp. $w' \in P_m$) if and only if $w \in P_m$ (resp. $w \in Q_m$). Since no other block is affected, we obtain an involution on the set of BWT images of generalized de Bruijn words, mapping P_m to Q_m. ◀

▶ **Lemma 25.** *Let $m > 0$. The set P_m (resp. Q_m) contains at least one totally unclustered word.*

Proof. From Lemma 24, we know that P_m and Q_m are both nonempty. Without loss of generality, let $w \in P_m$. We apply Lemma 19 iteratively to w, following the same procedure as in the proof of Theorem 3. At each step, we eliminate the rightmost tie (at some block $j \leq m-1$) until the word becomes totally unclustered. Crucially, this process does not affect the final block of the word, so the position $w_{3m-2} = 2$ remains unchanged. Hence, we obtain a totally unclustered word $w' \in P_m$. ◀

▶ **Lemma 26.** *Let $w \in \Sigma_3^{3m}$ be a totally unclustered alphabet-permutation power that is the BWT image of some necklace $[u]$, for some $m > 0$. If $w \in P_m$ (i.e., $w_{3m-2} = 2$), then $w' = w_0 \cdots w_{3m-3} w_{3m-1}$ is the BWT image of some necklace $[u']$. If $w \in Q_m$ (i.e., $w_{3m-3} = 2$), then $w'' = w_0 \cdots w_{3m-3} w_{3m-2} 2 w_{3m-1}$ is the BWT image of some necklace $[u'']$.*

Proof. Let us assume $w \in P_m$, and therefore $w_{3m-3} = 2$. Observe that the word w' of length $3m-1$ is obtained by deleting from w the 2 at position $3m-3$. As the last block of w is a permutation of Σ_3, we have $w_{3m-3} \neq w_{3m-1}$, so w' remains totally unclustered.

We now track the effect on the standard permutation. For all $i < 3m-2$, the relative order remains the same, so $\pi_{w'}(i) = \pi_w(i)$. Since the removed 2 was the last of its kind, we had $\pi_w(3m-2) = 3m-1$. The symbol at position $3m-1$ in w is now at $3m-2$ in w', so $\pi_{w'}(3m-2) = \pi_w(3m-1)$. All other values are preserved. Thus, $\pi_{w'}$ is obtained by deleting $3m-1$ from the cycle of π_w, which yields a $(3m-1)$-cycle. Therefore, $w' = BWT(u')$ for some necklace u' of length $3m-1$, and it is totally unclustered.

Now, assume that $w \in Q_m$, and therefore $w_{3m-1} = 2$. We define w'' as the word of length $3m+1$ obtained by inserting a 2 between positions $3m-2$ and $3m-1$ in w. This insertion is not adjacent to the existing 2 at position $3m-3$, so it increases the number of runs by 1, and w' remains totally unclustered.

Let us verify that $\pi_{w'}$, the standard permutation of w', is a cycle. Since the insertion occurs after the last 2 of w, all positions before $3m-1$ preserve their relative order: for all $i < 3m-1$, we have $\pi_{w'}(i) = \pi_w(i)$. The inserted 2 at position $3m-1$ is now the new last 2, so $\pi_{w'}(3m-1) = 3m$. The former position $3m-1$ (now shifted to $3m$) keeps its image: $\pi_{w'}(3m) = \pi_w(3m-1)$. All subsequent values are unchanged. Thus, $\pi_{w'}$ is obtained by inserting $3m$ just after $3m-1$ in the cycle of π_w. Since π_w is a $(3m)$-cycle, this operation yields a $(3m+1)$-cycle, so $w' = BWT(u')$ for some necklace u' of length $3m+1$, and is totally unclustered. ◀

Our main theorem now follows immediately from Lemma 25 and Lemma 26:

▶ **Theorem 27.** *For every integer $n > 0$, there exists a necklace $[u]$ of length n over the alphabet Σ_3 having totally unclustered BWT.*

▶ **Example 28.** Let $w = \texttt{201021120}$ as in Example 20. Note that $w \in P_3$. It has length 9, and removing ties yields $\hat{w} = \texttt{120102120} = BWT([\texttt{002212011}])$. Removing the last $\texttt{2}$ in $\hat{w}$ gives $\hat{w}' = \texttt{12010210}$, a totally unclustered word of length 8, which is the BWT of $[\hat{u}'] = [\texttt{00212011}]$, whose clustering ratio is $\frac{4}{3}$.

On the other hand, consider the word $\overline{w} = \texttt{201021201} = BWT([\texttt{001021122}]) \in Q_3$, which is totally unclustered. Hence, inserting a $\texttt{2}$ in penultimate position gives $\overline{w}'' = \texttt{2010212021}$, a totally unclustered word of length 10, which is the BWT of $[\overline{u}''] = [\texttt{0010211222}]$, whose clustering ratio is $\frac{5}{3}$.

Any ternary word $w \in \Sigma_3^*$ is a special case of a word over Σ_k, with $k > 3$, where only the letters $\texttt{0}$, $\texttt{1}$, and $\texttt{2}$ occur. Hence, the result of Theorem 27 can be easily extended to any alphabet of size larger than 3. It is natural to wonder whether the same result can be achieved if we require that $|w|_i \geq 1$ for all $i \in \Sigma_k$. We show the following:

▶ **Theorem 2.** *For every $k \geq 3$, and for every $n \geq k$, there exists a word u of length n over Σ_k such that every letter of Σ_k appears at least once and $BWT(u)$ is totally unclustered.*

Proof. Let $w \in \Sigma_3^n$ be any word having n runs that is a BWT image of some aperiodic necklace u, which exists by Theorem 27. We can then define a word w' as follows:

$$w'[\pi_w(n-i)] = \begin{cases} k-i & \text{if } 1 \leq i \leq k-3; \\ \texttt{2} & \text{if } i = k-2 \text{ and } k-3 \geq |w|_2|; \\ \texttt{1} & \text{if } i = k-1 \text{ and } k-3 \geq |w|_2 + |w|_1; \\ w[\pi_w(n-i)] & \text{otherwise.} \end{cases}$$

By construction, one can observe that $\pi_w = \pi_{w'}$, i.e., w' is the BWT image of some aperiodic necklace $[u']$ containing at least one occurrence of each character in Σ_k. Moreover, the word w' is obtained by replacing some one-letter runs in w with characters that occur only once in w', thus preserving the property of being totally unclustered, and the thesis follows. ◀

▶ **Example 29.** Consider the necklace $[u]$, where $u = \texttt{001021122}$, and its totally unclustered BWT $w = \texttt{201021201}$ with standard permutation $\pi_w = (0, 6, 8, 5, 4, 7, 2, 3, 1)$. The word $w' = \texttt{301041502} \in \Sigma_6^*$, which is obtained according to the method described in the proof of Theorem 2, is the BWT of the necklace $[u'] = [\texttt{001041253}]$. The clustering ratio of u' is $\frac{9}{8}$. A graphical representation is shown in Figure 5.

4 Lower Bound

For an alphabet of size at least 3, we can obtain a lower bound on the number of words having totally unclustered BWT, using the *generalized Euler's totient function*.

0	0	1	0	2	1	1	2	2
0	1	0	2	1	1	2	2	0
0	2	1	1	2	2	0	0	1
1	0	2	1	1	2	2	0	0
1	1	2	2	0	0	1	0	2
1	2	2	0	0	1	0	2	1
2	0	0	1	0	2	1	1	2
2	1	1	2	2	0	0	1	0
2	2	0	0	1	0	2	1	1

0	0	1	0	4	1	2	5	3
0	1	0	4	1	2	5	3	0
0	4	1	2	5	3	0	0	1
1	0	4	1	2	5	3	0	0
1	2	5	3	0	0	1	0	4
2	5	3	0	0	1	0	4	1
3	0	0	1	0	4	1	2	5
4	1	2	5	3	0	0	1	0
5	3	0	0	1	0	4	1	2

Figure 5 BWT tables for the necklaces $u = [001021122]$ (on the left) and $u' = [001041253]$ (on the right). The last column of each table corresponds to their respective BWT's $w = 201021201$ and $w' = 301041502$. We highlight with the same color the letters in the BWT tables of u and u' before and after the substitutions applied to w to obtain w', as described in the proof of Theorem 2.

Let p be a prime number. The generalized Euler's totient function counts the number of polynomials over $\mathbb{F}_p$ of degree smaller than m and coprime with $X^m - 1$. It can be computed using the formula

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

for $m > 1$, where $\lambda_p(m)$ is the largest power of p dividing m, and $\mathrm{ord}_p(d)$ is the multiplicative order of d modulo p.

▶ **Theorem 30.** *For every integers $n \geq k \geq 3$, writing $n = 3m + j$ with $j \in \{-1, 0, 1\}$, there exist at least $\Phi_3(m)/4m = \Omega(2^m/n) = \Omega(2^{\frac{n}{3}}/n)$ necklaces of length n over Σ_k having a totally unclustered BWT.*

Proof. The number of generalized de Bruijn words of length $n = 3m$ over Σ_3 is $2^{m-1} \cdot \frac{\Phi_3(m)}{m}$ [9].

By repeatedly applying Lemma 19 to any such word, we eventually obtain a totally unclustered BWT. Each application corresponds to choosing a subset $S \subseteq [0, \ldots, m-1]$ of blocks to untie. Since there are 2^m possible subsets, every totally unclustered word can arise as the image of at most 2^m generalized de Bruijn words.

Applying this argument to the whole family of generalized de Bruijn words yields at least $\Phi_3(m)/2m$ totally unclustered words of length $n = 3m$. Applying it instead to the sets P_m and Q_m, we obtain at least $\Phi_3(m)/4m$ totally unclustered words in each of them, since Lemma 24 gives

$$|P_m| = |Q_m| = 2^{m-1} \cdot \frac{\Phi_3(m)}{2m}.$$

Finally, by Lemma 26, each totally unclustered word in P_m (resp. Q_m) is injectively mapped to a necklace of length $3m - 1$ (resp. $3m + 1$) with a totally unclustered BWT. Hence, there are at least $\Phi_3(m)/4m$ such necklaces of length $3m - 1$ and the same number of length $3m + 1$.

From equation 1, and observing that $\mathrm{ord}_3(d) \geq 1$ for all d, we have

$$\sum_{d|(m/\lambda_3(m))} \frac{\phi(d)}{\mathrm{ord}_p(d)} \leq \sum_{d|(m/\lambda_3(m))} \phi(d) = \frac{m}{\lambda_3(m)}.$$

Therefore,

$$\Phi_3(m) \geq 3^m\left(1-\tfrac{1}{3}\right)^{m/\lambda_p(m)} \geq 3^m\left(1-\tfrac{1}{3}\right)^m \geq 2^m,$$

and since $\Sigma_3^n \subseteq \Sigma_k^n$, the thesis follows. ◀

5 Special Case related to Artin's Conjecture

We showed that necklaces with totally unclustered BWT exist for every length when the alphabet has size at least 3, contrasting with the binary case (Theorem 1). In particular, this case is related to a famous conjecture of Emil Artin (see [17] for a more detailed presentation).

An integer q is a *primitive root modulo* p if q generates multiplicatively the group $\mathbb{Z}_p^\times$, namely if $\{q^i \mod p \mid i \in \mathbb{Z}\} = \{1, \ldots, p-1\}$.

▶ **Conjecture 31.** *Let q be an integer that is not a square number and not -1. Then, q is a primitive root modulo p for infinitely many primes p.*

This happens because, in the binary case, a word that is a BWT image and is totally unclustered is necessarily of the form $(10)^n$. Reciprocally, the word $(10)^n$ is a BWT image if and only if $2n+1$ is an odd prime and 2 generates the cyclic group $\mathbb{Z}_{2n+1}^\times$, as stated in Theorem 1. This is no longer the case if the alphabet contains at least 3 letters, since more words have maximal RLE. However, one can still find a trace of this phenomenon in the larger alphabet case:

▶ **Theorem 32.** *Let $k \geq 2$, and $v = (k-1)\cdots\mathtt{0}$ be the word obtained by concatenating each letter of Σ_k in decreasing lexicographic order. For every integer m, the word v^m is a BWT image if and only if $km+1$ is a prime and k is a primitive root modulo $km+1$.*

Proof. The proof is essentially the same as in the binary case [24]. By relabeling and considering the standard permutation π_{v^m} on the set $\{1, \ldots, km\}$, one can express $\pi_{v^m}(i) = ki \mod km+1$, hence $\pi_{v^m}^i(1) = k^i \mod km+1$. This is a single cycle if and only if $\{k^i \mod km+1 \mid i \in \mathbb{Z}\} = \{1, \ldots, km\}$. Since k and $km+1$ are coprime, the left-hand set is contained in $\mathbb{Z}_{km+1}^\times$, and since $\mathbb{Z}_{km+1}^\times \subseteq \{1, \ldots, km\}$, the three sets are equal. In particular, $|\mathbb{Z}_{km+1}^\times| = km$, which holds if and only if $km+1$ is prime. Thus $km+1$ is prime and k is a primitive root modulo $km+1$. ◀

▶ Remark 33. Note that, when v^m is a BWT image, the necklace u such that $BWT(u) = v^m$ is, by definition, a generalized de Bruijn word (see also Example 14 for the case $k=3$ and $m=6$). However, it is never an ordinary de Bruijn word (equivalently, an integer m such that $km = k^d$ for some integer d never satisfies the condition of Theorem 32). This was shown in [24] for the case $k=2$. For $k>2$, observe that $k^d \equiv -1 \mod k^d+1$, so $k^{2d} \equiv 1 \mod k^d+1$. This means that the order of $k \mod k^d+1$ must divide $2d$. But $2d \leq 3^d \leq k^d$ for every $d \geq 1$, so k cannot have order k^d, and k^d+1 cannot have k as a primitive root.

6 Conclusions and Future Work

In this work, we proved that for every alphabet of size at least three and for every length $n > 0$, a necklace (and hence a class of words) whose BWT is totally unclustered can be constructed. This shows that the worst-case behavior of the BWT with respect to clustering is not only possible but also unavoidable for all lengths once the alphabet has at least three symbols. Our approach is based on generalized de Bruijn words and their interpretation as Hamiltonian cycles in generalized de Bruijn graphs, from which we also derived a lower bound on the number of such words.

These results open new research directions. From a combinatorial perspective, it would be interesting to characterize all necklaces with totally unclustered BWT (in particular, those that are not generalized de Bruijn words), and to derive tighter lower bounds on their number.

Another natural research direction is the study of how other compression-based repetitiveness measures behave on these extremal cases.

Finally, although the existence of infinitely many binary words with totally unclustered BWT is still unproven, it has been shown that if a binary word u has unclustered BWT, then the clustering ratio can only take the values 2 or $2 - 2/|\operatorname{RLE}(u')|$, where u' is the lexicographically smallest rotation of u [24, Propositions 4.5 and 4.7]. For arbitrary alphabets, following Example 7 and Remark 23, we leave open the problem of characterizing words that have both a totally unclustered BWT and a clustering ratio of 2.

References

1 Tooru Akagi, Mitsuru Funakoshi, and Shunsuke Inenaga. Sensitivity of string compressors and repetitiveness measures. *Inf. Comput.*, 291:104999, 2023. doi:10.1016/J.IC.2022.104999.

2 Yu Hin Au. Generalized de Bruijn words for primitive words and powers. *Discrete Mathematics*, 338(12):2320–2331, 2015. doi:10.1016/j.disc.2015.05.025.

3 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. doi:10.1007/978-1-4757-0602-4_6.

4 Michael Burrows and David J. Wheeler. A Block-sorting Lossless Data Compression Algorithm. *SRS Research Report*, 124, 1994.

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

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

7 Sébastien Ferenczi and Luca Q. Zamboni. Clustering Words and Interval Exchanges. *Journal of Integer Sequences*, 16(2):Article 13.2.1, 2013. URL: https://cs.uwaterloo.ca/journals/JIS/VOL16/Zamboni/zamboni2.pdf.

8 Paolo Ferragina and Giovanni Manzini. Indexing compressed text. *J. ACM*, 52(4):552–581, 2005. doi:10.1145/1082036.1082039.

9 Gabriele Fici and Estéban Gabory. Generalized De Bruijn Words, Invertible Necklaces, and the Burrows-Wheeler Transform. In Paweł Gawrychowski, Filip Mazowiecki, and Michał Skrzypczak, editors, *50th International Symposium on Mathematical Foundations of Computer Science (MFCS 2025)*, volume 345 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 48:1–48:18, Dagstuhl, Germany, 2025. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.MFCS.2025.48.

10 Gabriele Fici, Sabrina Mantaci, Antonio Restivo, Giuseppe Romana, Giovanna Rosone, and Marinella Sciortino. BWT and combinatorics on words. In *The Expanding World of Compressed Data*, volume 131 of *OASIcs*, pages 1:1–1:23. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025. doi:10.4230/OASIcs.MANZINI.1.

11 Gabriele Fici, Giuseppe Romana, Marinella Sciortino, and Cristian Urbina. On the Impact of Morphisms on BWT-Runs. In *CPM 2023*, volume 259 of *LIPIcs*, pages 10:1–10:18. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi:10.4230/LIPIcs.CPM.2023.10.

12 Gabriele Fici, Giuseppe Romana, Marinella Sciortino, and Cristian Urbina. Morphisms and BWT-Run Sensitivity. In *MFCS*, volume 345 of *LIPIcs*, pages 49:1–49:18. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025. doi:10.4230/LIPIcs.MFCS.2025.49.

13 Travis Gagie, Gonzalo Navarro, and Nicola Prezza. Fully functional suffix trees and optimal text searching in BWT-runs bounded space. *J. ACM*, 67(1):2:1–2:54, 2020. doi:10.1145/3375890.

14 Sara Giuliani, Shunsuke Inenaga, Zsuzsanna Lipták, Nicola Prezza, Marinella Sciortino, and Anna Toffanello. Novel Results on the Number of Runs of the Burrows-Wheeler-Transform. In *SOFSEM*, volume 12607 of *Lecture Notes in Computer Science*, pages 249–262. Springer, 2021. doi:10.1007/978-3-030-67731-2_18.

15 Sara Giuliani, Shunsuke Inenaga, Zsuzsanna Lipták, Giuseppe Romana, Marinella Sciortino, and Cristian Urbina. Bit Catastrophes for the Burrows-Wheeler Transform. *Theory Comput. Syst.*, 69(2):19, 2025. doi:10.1007/s00224-024-10212-9.

16 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.

17 Christopher Hooley. On Artin's conjecture. *Journal für die reine und angewandte Mathematik*, 1967(225):209–220, 1967. doi:10.1515/crll.1967.225.209.

18 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.

19 Dominik Kempa and Tomasz Kociumaka. Resolution of the Burrows-Wheeler transform conjecture. *Commun. ACM*, 65(6):91–98, 2022. doi:10.1145/3531445.

20 Mélodie Lapointe and Christophe Reutenauer. Characterizations of perfectly clustering words. *Electron. J. Comb.*, 32(3), 2025. doi:10.37236/12851.

21 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.

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

23 Sabrina Mantaci, Antonio Restivo, Giovanna Rosone, and Marinella Sciortino. An extension of the Burrows-Wheeler Transform. *Theor. Comput. Sci.*, 387(3):298–312, 2007. doi:10.1016/J.TCS.2007.07.014.

24 Sabrina Mantaci, Antonio Restivo, Giovanna Rosone, Marinella Sciortino, and Luca Versari. Measuring the clustering effect of BWT via RLE. *Theoretical Computer Science*, 698:79–87, October 2017. doi:10.1016/j.tcs.2017.07.015.

25 Sabrina Mantaci, Antonio Restivo, and Marinella Sciortino. Burrows-Wheeler transform and Sturmian words. *Information Processing Letters*, 86:241–246, 2003. doi:10.1016/S0020-0190(02)00512-4.

26 Giovanni Manzini. An analysis of the Burrows-Wheeler transform. *J. ACM*, 48(3):407–430, 2001. doi:10.1145/382780.382782.

27 Gonzalo Navarro. Indexing Highly Repetitive String Collections, Part I: Repetitiveness Measures. *ACM Comput. Surv.*, 54(2):29:1–29:31, 2022. doi:10.1145/3434399.

28 Gonzalo Navarro. Technical perspective: The compression power of the BWT. *Commun. ACM*, 65(6):90, 2022. doi:10.1145/3531443.

29 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.

30 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.

31 Bruce Sagan. *The symmetric group: representations, combinatorial algorithms, and symmetric functions*, volume 203. Springer Science & Business Media, 2001.