Reconstructing Words from Right-Bounded-Block Words

Pamela Fleischmann[1], Marie Lejeune[2⋆], Florin Manea[3⋆⋆], Dirk Nowotka[1], and Michel Rigo[2]

[1] Kiel University, Germany {fpa,dn}@informatik.uni-kiel.de
[2] University of Liège, Belgium {m.lejeune,m.rigo}@uliege.be
[3] University of Göttingen, Germany florin.manea@informatik.uni-goettingen.de

Abstract. A reconstruction problem of words from scattered factors asks for the minimal information, like multisets of scattered factors of a given length or the number of occurrences of scattered factors from a given set, necessary to uniquely determine a word. We show that a word $w \in \{\mathtt{a}, \mathtt{b}\}^*$ can be reconstructed from the number of occurrences of at most $\min(|w|_\mathtt{a}, |w|_\mathtt{b}) + 1$ scattered factors of the form $\mathtt{a}^i\mathtt{b}$, where $|w|_a$ is the number of occurrences of the letter a in w. Moreover, we generalize the result to alphabets of the form $\{1, \ldots, q\}$ by showing that at most $\sum_{i=1}^{q-1} |w|_i \, (q - i + 1)$ scattered factors suffices to reconstruct w. Both results improve on the upper bounds known so far. Complexity time bounds on reconstruction algorithms are also considered here.

1 Introduction

The general scheme for a so-called *reconstruction problem* is the following one: given a sufficient amount of information about substructures of a hidden discrete structure, can one uniquely determine this structure? In particular, what are the fragments about the structure needed to recover it all. For instance, a square matrix of size at least 5 can be reconstructed from its principal minors given in any order [20].

In graph theory, given some subgraphs of a graph (these subgraphs may share some common vertices and edges), can one uniquely rebuild the original graph? Given a finite undirected graph $G = (V, E)$ with n vertices, consider the multiset made of the n induced subgraphs of G obtained by deleting exactly one vertex from G. In particular, one knows how many isomorphic subgraphs of a given class appear. Two graphs leading to the same multiset (generally called a *deck*) are said to be *hypomorphic*. A conjecture due to Kelly and Ulam states that two hypomorphic graphs with at least three vertices are isomorphic [14, 21]. A similar conjecture in terms of edge-deleted subgraphs has been proposed by Harary [11]. These conjectures are known to hold true for several families of graphs.

⋆ Supported by a FNRS fellowship.
⋆⋆ Supported by the DFG grant MA 5725/2-1.

A *finite word*, i.e., a finite sequence of letters of some given alphabet, can be seen as an edge- or vertex-labeled linear tree. So variants of the graph reconstruction problem can be considered and are of independent interest. Participants of the Oberwolfach meeting on Combinatorics on Words in 2010 [2] gave a list of 18 important open problems in the field. Amongst them, the twelfth problem is stated as *reconstruction from subwords of given length*. In the following statement and all along the paper, a *subword* of a word is understood as a subsequence of not necessarily contiguous letters from this word, i.e., subwords can be obtained by deleting letters from the given word. To highlight this latter property, they are often called *scattered subwords* or *scattered factors*, which is the notion we are going to use.

Definition 1. *Let k, n be natural numbers. Words of length n over a given alphabet are said to be* k-reconstructible *whenever the multiset of scattered factors of length k (or* k-deck*) uniquely determines any word of length n.*

Notice that the definition requires multisets to store the information how often a scattered factor occurs in the words. For instance, the scattered factor `ba` occurs three times in `baba` which provides more information for the reconstruction than the mere fact that `ba` is a scattered factor.

The challenge is to determine the function $f(n) = k$ where k is the least integer for which words of length n are k-reconstructible. This problem has been studied by several authors and one of the first trace goes back to 1973 [13]. Results in that direction have been obtained by M.-P. Schützenberger (with the so-called *Schützenberger's Guessing game*) and L. Simon [25]. They show that words of length n sharing the same multiset of scattered factors of length up to $\lfloor n/2 \rfloor + 1$ are the same. Consequently, words of length n are $(\lfloor n/2 \rfloor + 1)$-reconstructible. In [15], this upper bound has been improved: Krasikov and Roditty have shown that words of length n are k-reconstructible for $k \geq \lfloor 16\sqrt{n}/7 \rfloor + 5$. On the other hand Dudik and Schulmann [6] provide a lower bound: if words of length n are k-reconstructible, then $k \geq 3^{(\sqrt{2/3} - o(1)) \log_3^{1/2} n}$. Bounds were also considered in [19]. Algorithmic complexity of the reconstruction problem is discussed, for instance, in [5]. Note that the different types of reconstruction problems have application in philogenetic networks, see, e.g., [12], or in the context of molecular genetics [7] and coding theory [16].

Another motivation, close to combinatorics on words, stems from the study of k-binomial equivalence of finite words and k-binomial complexity of infinite words (see [23] for more details). Given two words of the same length, they are k-binomially equivalent if they have the same multiset of scattered factors of length k, also known as *k-spectrum* ([1], [18], [24]). Given two words x and y of the same length, one can address the following problem: decide whether or not x and y are k-binomially equivalent? A polynomial time decision algorithm based on automata and a probabilistic algorithm have been addressed in [10]. A variation of our work would be to find, given k and n, a minimal set of scattered factors for which the knowledge of the number of occurrences in x and y permits to decide k-binomial equivalence.

Over an alphabet of size q, there are q^k pairwise distinct length-k factors. If we relax the requirement of only considering scattered factors of the same length, another interesting question is to look for a minimal (in terms of cardinality) multiset of scattered factors to reconstruct entirely a word. Let the *binomial coefficient* $\binom{u}{x}$ be the number of occurrences of x as a scattered factor of u. The general problem addressed in this paper is therefore the following one.

Problem 2. Let Σ be a given alphabet and n a natural number. We want to reconstruct an unknown word $w \in \Sigma^n$. To that aim, we are allowed to ask questions of the type "What is the value of $\binom{w}{u_i}$?". Based on the answers to questions related to $\binom{w}{u_1}, \ldots, \binom{w}{u_i}$, we can decide which will be the next question (i.e. decide which word will be u_{i+1}). We want to have the shortest sequence $(u_1, \ldots, u_k)$ uniquely determining w by knowing the values of $\binom{w}{u_1}, \ldots, \binom{w}{u_k}$.

In this new context, we naturally look for a value of k less than the upper bound for k-reconstructibility.

In this paper, we firstly recall the use of Lyndon words in the context of reconstructibility. A word w over a totally ordered alphabet is called *Lyndon word* if it is the lexicographically smallest amongst all its rotations, i.e., $w = xy$ is smaller than yx for all non trivial factorisations $w = xy$. Every binomial coefficient $\binom{w}{x}$ for arbitrary words w and x over the same alphabet can be deduced from the values of the coefficients $\binom{w}{u}$ for Lyndon words u that are lexicographically less than or equal to x. This result is presented in Section 2 along with the basic definitions. We consider an alphabet equipped with a total order on the letters. Words of the form $\mathtt{a}^n\mathtt{b}$ with letters $\mathtt{a} < \mathtt{b}$ and a natural number n are a special form of Lyndon words, the so-called *right-bounded-block* words.

We consider the reconstruction problem from the information given by the occurrences of right-bounded-block words as scattered factors of a word of length n. In Section 3 we show how to reconstruct a word uniquely from $m + 1$ binomial coefficients of right-bounded-block words where m is the minimum number of occurrences of $\mathtt{a}$ and $\mathtt{b}$ in the word. We also prove that this is less than the upper bound given in [15]. In Section 4 we reduce the problem for arbitrary finite alphabets $\{1, \ldots, q\}$ to the binary case. Here we show that at most $\sum_{i=1}^{q-1} |w|_i \, (q - i + 1) \leq q|w|$ binomial coefficients suffice to uniquely reconstruct w with $|w|_i$ being the number of occurrences of letter i in w. Again, we compare this bound to the best known one for the classical reconstruction problem (from words of a given length). In the last section of the paper we also propose several results of algorithmic nature regarding the efficient reconstruction of words from given scattered factors.

Due to space restrictions some proofs (marked with $*$) can be found in [9].

2 Preliminaries

Let $\mathbb{N}$ be the set of natural numbers, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, and let $\mathbb{N}_{\geq k}$ be the set of all natural numbers greater than or equal to k. Let $[n]$ denote the set $\{1, \ldots, n\}$ and $[n]_0 = [n] \cup \{0\}$ for an $n \in \mathbb{N}$.

An *alphabet* $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \dots\}$ is a finite set of letters and a *word* is a finite sequence of letters. We let Σ^* denote the set of all finite words over Σ. The *empty word* is denoted by ε and Σ^+ is the free semigroup $\Sigma^* \backslash \{\varepsilon\}$. The length of a word w is denoted by $|w|$. Let $\Sigma^{\leq k} := \{w \in \Sigma^* \mid |w| \leq k\}$ and Σ^k be the set of all words of length exactly $k \in \mathbb{N}$. The number of occurrences of a letter $\mathtt{a} \in \Sigma$ in a word $w \in \Sigma^*$ is denoted by $|w|_{\mathtt{a}}$. The i^{th} letter of a word w is given by $w[i]$ for $i \in [|w|]$. The powers of $w \in \Sigma^*$ are defined recursively by $w^0 = \varepsilon$, $w^n = ww^{n-1}$ for $n \in \mathbb{N}$. A word $u \in \Sigma^*$ is a *factor* of $w \in \Sigma^*$, if $w = xuy$ holds for some words $x, y \in \Sigma^*$. Moreover, u is a *prefix* of w if $x = \varepsilon$ holds and a *suffix* if $y = \varepsilon$ holds. The factor of w from the i^{th} to the j^{th} letter will be denoted by $w[i..j]$ for $1 \leq i \leq j \leq |w|$. Two words $u, v \in \Sigma^*$ are called *conjugates or rotations* of each other if there exist $x, y \in \Sigma^*$ with $u = xy$ and $v = yx$. Additional basic information about combinatorics on words can be found in [17].

Definition 3. *Let $<$ be a total ordering on Σ. A word $w \in \Sigma^*$ is called* right-bounded-block word *if there exist $\mathtt{x}, \mathtt{y} \in \Sigma$ with $\mathtt{x} < \mathtt{y}$ and $\ell \in \mathbb{N}_0$ with $w = \mathtt{x}^\ell \mathtt{y}$.*

Definition 4. *A word $u = \mathtt{a}_1 \cdots \mathtt{a}_n \in \Sigma^n$, for $n \in \mathbb{N}$, is a* scattered factor *of a word $w \in \Sigma^+$ if there exist $v_0, \dots, v_n \in \Sigma^*$ with $w = v_0 \mathtt{a}_1 v_1 \cdots v_{n-1} \mathtt{a}_n v_n$. For words $w, u \in \Sigma^*$, define $\binom{w}{u}$ as the number of occurrences of u as a scattered factor of w.*

Remark 5. Notice that $|w|_{\mathtt{x}} = \binom{w}{\mathtt{x}}$ for all $\mathtt{x} \in \Sigma$.

The following definition addresses Problem 2.

Definition 6. *A word $w \in \Sigma^n$ is called* uniquely reconstructible/determined *by the set $S \subset \Sigma^*$ if for all words $v \in \Sigma^n \backslash \{w\}$ there exists a word $u \in S$ with $\binom{w}{u} \neq \binom{v}{u}$.*

Consider $S = \{\mathtt{ab}, \mathtt{ba}\}$. Then $w = \mathtt{abba}$ is not uniquely reconstructible by S since $\left[\binom{w}{\mathtt{ab}}, \binom{w}{\mathtt{ba}}\right] = [2, 2]$ is also the 2-vector of binomial coefficients of $\mathtt{baab}$. On the other hand $S = \{\mathtt{a}, \mathtt{ab}, \mathtt{ab}^2\}$ reconstructs w uniquely. The following remark gives immediate results for binary alphabets.

Remark 7. Let $\Sigma = \{\mathtt{a}, \mathtt{b}\}$ and $w \in \Sigma^n$. If $|w|_{\mathtt{a}} \in \{0, n\}$ then w contains either only $\mathtt{b}$ or $\mathtt{a}$ and by the given length n of w, w is uniquely determined by $S = \{\mathtt{a}\}$. This fact is in particular an equivalence: $w \in \Sigma^n$ can be uniquely determined by $\{\mathtt{a}\}$ iff $|w|_{\mathtt{a}} \in \{0, n\}$. If $|w|_{\mathtt{a}} \in \{1, n-1\}$, w is not uniquely determined by $\{\mathtt{a}\}$ as witnessed by $\mathtt{ab}$ and $\mathtt{ba}$ for $n = 2$. It is immediately clear that the additional information $\binom{w}{\mathtt{ab}}$ leads to unique determinism of w.

Lyndon words play an important role regarding the reconstruction problem. As shown in [22] only scattered factors which are Lyndon words are necessary to determine a word uniquely, i.e., S can always be assumed to be a set of Lyndon words.

Definition 8. *Let $<$ be a total ordering on Σ. A word $w \in \Sigma^*$ is a* Lyndon word *iff for all $u, v \in \Sigma^+$ with $w = uv$, we have $w <_{lex} vu$ where $<_{lex}$ is the lexicographical ordering on words induced by $<$.*

Proposition 9 ([22]). *Let w and u be two words. The binomial coefficient $\binom{w}{u}$ can be computed using only binomial coefficients of the type $\binom{w}{v}$ where v is a Lyndon word of length up to $|u|$ such that $v \leq_{lex} u$.*

To obtain a formula to compute the binomial coefficient $\binom{w}{u}$ for $w, u \in \Sigma^*$ by binomial coefficients $\binom{w}{v_i}$ for Lyndon words $v_1, \ldots, v_k$ with $v_i \in \Sigma^{\leq |u|}$, $i \in [k]$, and $k \in \mathbb{N}$ the definitions of shuffle and infiltration are necessary [17].

Definition 10. *Let $n_1, n_2 \in \mathbb{N}$, $u_1 \in \Sigma^{n_1}$, and $u_2 \in \Sigma^{n_2}$. Set $n = n_1 + n_2$. The* shuffle *of u_1 and u_2 is the polynomial $u_1 \sqcup\!\sqcup u_2 = \sum_{I_1, I_2} w(I_1, I_2)$ where the sum has to be taken over all pairs (I_1, I_2) of sets that are partitions of $[n]$ such that $|I_1| = n_1$ and $|I_2| = n_2$. If $I_1 = \{i_{1,1} < \ldots < i_{1,n_1}\}$ and $I_2 = \{i_{2,1} < \ldots < i_{2,n_2}\}$, then the word $w(I_1, I_2)$ is defined such that $w[i_{1,1}]w[i_{1,2}] \cdots w[i_{1,n_1}] = u_1$ and $w[i_{2,1}]w[i_{2,2}] \cdots w[i_{2,n_2}] = u_2$ hold.*

The infiltration is a variant of the shuffle in which equal letters are merged.

Definition 11. *Let $n_1, n_2 \in \mathbb{N}$, $u_1 \in \Sigma^{n_1}$, and $u_2 \in \Sigma^{n_2}$. Set $n = n_1 + n_2$. The* infiltration *of u_1 and u_2 is the polynomial $u_1 \downarrow u_2 = \sum_{I_1, I_2} w(I_1, I_2)$, where the sum has to be taken over all pairs (I_1, I_2) of sets of cardinality n_1 and n_2 respectively, for which the union is equal to the set $[n']$ for some $n' \leq n$. Words $w(I_1, I_2)$ are defined as in the previous definition. Note that some $w(I_1, I_2)$ are not well defined if $i_{1,j} = i_{2,k}$ but $u_1[j] \neq u_2[k]$. In that case they do not appear in the previous sum.*

Considering for instance $u_1 = \texttt{aba}$ and $u_2 = \texttt{ab}$ gives the polynomials

$$u_1 \sqcup\!\sqcup u_2 = 2\texttt{ababa} + 4\texttt{aabba} + 2\texttt{aabab} + 2\texttt{abaab},$$
$$u_1 \downarrow u_2 = \texttt{aba} \sqcup\!\sqcup \texttt{ab} + \texttt{aba} + 2\texttt{abba} + 2\texttt{aaba} + 2\texttt{abab}.$$

Based on Definitions 10 and 11, we are able to give a formula to compute a binomial coefficient from the ones making use of Lyndon words. This formula is given implicitely in [22, Theorem 6.4]: Let $u \in \Sigma^*$ be a non-Lyndon word. By [22, Corollary 6.2] there exist non-empty words $x, y \in \Sigma^*$ and with $u = xy$ and such that every word appearing in the polynomial $x \sqcup\!\sqcup y$ is lexicographically less than or equal to u. Then, for all word $w \in \Sigma^*$, we have

$$\binom{w}{u} = \frac{1}{(x \sqcup\!\sqcup y, u)} \left[\binom{w}{x}\binom{w}{y} - \sum_{v \in \Sigma^*, v \neq u} (x \downarrow y, v) \binom{w}{v} \right],$$

where (P, v) is a notation giving the coefficient of the word v in the polynomial P. One may apply recursively this formula until only Lyndon factors are considered. Some examples can be found in the appendix.

3 Reconstruction from Binary Right-Bounded-Block Words

In this section we present a method to reconstruct a binary word uniquely from binomial coefficients of right-bounded-block words. Let $n \in \mathbb{N}$ be a natural number and $w \in \{\mathtt{a}, \mathtt{b}\}^n$ a word. Since the word length n is assumed to be known, $|w|_\mathtt{a}$ is known if $|w|_\mathtt{b}$ is given and vice versa. Set for abbreviation $k_u = \binom{w}{u}$ for $u \in \Sigma^*$. Moreover we assume w.l.o.g. $k_\mathtt{a} \leq k_\mathtt{b}$ and that $k_\mathtt{a}$ is known (otherwise substitute each $\mathtt{a}$ by $\mathtt{b}$ and each $\mathtt{b}$ by $\mathtt{a}$, apply the following reconstruction method and revert the substitution). This implies that w is of the form

$$\mathtt{b}^{s_1}\mathtt{a}\mathtt{b}^{s_2}\dots\mathtt{b}^{s_{k_\mathtt{a}}}\mathtt{a}\mathtt{b}^{s_{k_\mathtt{a}+1}} \tag{1}$$

for $s_i \in \mathbb{N}_0$ and $i \in [|w|_\mathtt{a} + 1]$ with $\sum_{i\in[k_\mathtt{a}+1]} s_i = n - k_\mathtt{a} = k_\mathtt{b}$ and thus we get for $\ell \in [k_\mathtt{a}]_0$

$$k_{\mathtt{a}^\ell\mathtt{b}} = \binom{w}{\mathtt{a}^\ell\mathtt{b}} = \sum_{i=\ell+1}^{k_\mathtt{a}+1} \binom{i-1}{\ell} s_i. \tag{2}$$

Remark 12. Notice that for fixed $\ell \in [k_\mathtt{a}]_0$ and $c_i = \binom{i-1}{\ell}$ for $i \in [k_\mathtt{a}+1]\backslash[\ell]$, we have $c_i < c_{i+1}$ and especially $c_{\ell+1} = 1$ and $c_{\ell+2} = \ell + 1$.

Equation (2) shows that reconstructing a word uniquely from binomial coefficients of right-bounded-block words equates to solve a system of Diophantine equations. The knowledge of $k_\mathtt{b}, \dots, k_{\mathtt{a}^\ell\mathtt{b}}$ provides $\ell + 1$ equations. If the equation of $k_{\mathtt{a}^\ell\mathtt{b}}$ has a unique solution for $\{s_{\ell+1}, \dots, s_{k_\mathtt{a}+1}\}$ (in this case we say, by language abuse, that $k_{\mathtt{a}^\ell\mathtt{b}}$ is *unique*), then the system in row echelon form has a unique solution and thus the binary word is uniquely reconstructible. Notice that $k_{\mathtt{a}^{k_\mathtt{a}}\mathtt{b}}$ is always unique since $k_{\mathtt{a}^{k_\mathtt{a}}\mathtt{b}} = s_{k_\mathtt{a}+1}$.

Consider $n = 10$ and $k_\mathtt{a} = 4$. This leads to $w = \mathtt{b}^{s_1}\mathtt{a}\mathtt{b}^{s_2}\mathtt{a}\mathtt{b}^{s_3}\mathtt{a}\mathtt{b}^{s_4}\mathtt{a}\mathtt{b}^{s_5}$ with $\sum_{i\in[5]} s_i = 6$. Given $k_{\mathtt{a}\mathtt{b}} = 4$ we get $4 = s_2 + 2s_3 + 3s_4 + 4s_5$. The s_i are not uniquely determined. If $k_{\mathtt{a}^2\mathtt{b}} = 2$ is also given, we obtain the equation $2 = s_3 + 3s_4 + 6s_5$ and thus $s_3 = 2$ and $s_4 = s_5 = 0$ is the only solution. Substituting these results in the previous equation leads to $s_2 = 0$ and since we only have six $\mathtt{b}$, we get $s_1 = 4$. Hence $w = \mathtt{b}^4\mathtt{a}^2\mathtt{b}^2\mathtt{a}^2$ is uniquely reconstructed by $S = \{\mathtt{a}, \mathtt{a}\mathtt{b}, \mathtt{a}^2\mathtt{b}\}$.

The following definition captures all solutions for the equation defined by $k_{\mathtt{a}^\ell\mathtt{b}}$ for $\ell \in [k_\mathtt{a}]_0$.

Definition 13. *Set* $M(k_{\mathtt{a}^\ell\mathtt{b}}) = \{(r_{\ell+1}, \dots, r_{k_\mathtt{a}+1}) \mid k_{\mathtt{a}^\ell\mathtt{b}} = \sum_{i=\ell+1}^{k_\mathtt{a}+1} \binom{i-1}{\ell} r_i\}$ *for fixed* $\ell \in [k_\mathtt{a}]_0$. *We call* $k_{\mathtt{a}^\ell\mathtt{b}}$ unique *if* $|M(k_{\mathtt{a}^\ell\mathtt{b}})| = 1$.

By Remark 12 the coefficients of each equation of the form (2) are strictly increasing. The next lemma provides the range each $k_{\mathtt{a}^\ell\mathtt{b}}$ may take under the constraint $\sum_{i=1}^{k_\mathtt{a}+1} s_i = n - k_\mathtt{a}$.

Lemma 14. *Let* $n \in \mathbb{N}$, $k \in [n]_0$, $j \in [k+1]$ *and* $c_1, \dots, c_{k+1}, s_1, \dots, s_{k+1} \in \mathbb{N}_0$ *with* $c_i < c_{i+1}$, *for* $i \in [k]$, *and* $\sum_{i=1}^{k+1} s_i = n - k$. *The sum* $\sum_{i=j}^{k+1} c_i s_i$ *is maximal iff* $s_{k+1} = n - k$ *(and consequently* $s_i = 0$ *for all* $i \in [k]$*).*

Proof. The case $k = 0$ is trivial. Consider the case $n = k$, i.e., $\sum_{i=1}^{k+1} s_i = 0$. This implies immediately $s_i = 0$ for all $i \in [k+1]$ and the equivalence holds. Assume for the rest of the proof $k < n$. If $s_{k+1} = n - k$, then $s_i = 0$ for all $i \leq k$ and $\sum_{i=j}^{k+1} c_i s_i = c_{k+1}(n-k)$. Let us assume that the maximal value for $\sum_{i=j}^{k+1} c_i s_i$ can be obtained in another way and that there exist $s'_1, \ldots, s'_{k+1} \in \mathbb{N}_0$, $\ell \in [n-k]$ such that $\sum_{i=1}^{k+1} s'_i = n - k$ and $s'_{k+1} = n - k - \ell$. Thus

$$c_{k+1}(n-k) \leq \sum_{i=j}^{k+1} c_i s'_i = \left(\sum_{i=j}^{k} c_i s'_i \right) + c_{k+1}(n-k-\ell).$$

This implies $\sum_{i=j}^{k} c_i s'_i \geq c_{k+1}\ell$. Since the coefficients are strictly increasing we get $\sum_{i=j}^{k} c_i s'_i \leq c_k \sum_{i=j}^{k} s'_i < c_{k+1}\ell$, hence the contradiction. □

Corollary 15. *Let* $k_{\mathtt{a}} \in [n]_0$, $\ell \in [k_{\mathtt{a}}]_0$, *and* $s_1, \ldots, s_{k_{\mathtt{a}}+1} \in \mathbb{N}_0$ *with* $\sum_{i=1}^{k_{\mathtt{a}}+1} s_i = n - k_{\mathtt{a}}$. *Then* $\binom{w}{\mathtt{a}^\ell \mathtt{b}} \in \left[\binom{k_{\mathtt{a}}}{\ell}(n - k_{\mathtt{a}}) \right]_0$.

Proof. It follows directly from Equation (2) and Lemma 14. □

The following lemma shows some cases in which $k_{\mathtt{a}^\ell \mathtt{b}}$ is unique.

Lemma 16. *Let* $k_{\mathtt{a}} \in [n]$, $\ell \in [k_{\mathtt{a}}]_0$ *and* $s_1, \ldots, s_{k_{\mathtt{a}}+1} \in \mathbb{N}_0$ *with* $\sum_{i=1}^{k_{\mathtt{a}}+1} s_i = n - k_{\mathtt{a}}$. *If* $k_{\mathtt{a}^\ell \mathtt{b}} \in [\ell]_0 \cup \{\binom{k_{\mathtt{a}}}{\ell}(n - k_{\mathtt{a}})\}$ *or* $k_{\mathtt{a}^\ell \mathtt{b}} = \binom{k_{\mathtt{a}}-1}{\ell} r + \binom{k_{\mathtt{a}}}{\ell}(n - k_{\mathtt{a}} - r)$ *for* $r \in [k_{\mathtt{b}}]_0$ *then* $k_{\mathtt{a}^\ell \mathtt{b}}$ *is unique.*

Proof. Consider firstly $k_{\mathtt{a}^\ell \mathtt{b}} \in [\ell]_0$. By Remark 12 we have $c_{\ell+1} = 1$ and $c_{\ell+2} = \ell + 1$. By $c_i < c_{i+1}$ we obtain immediately $s_i = 0$ for $i \in [k_{\mathtt{a}} + 1] \backslash [\ell + 1]$. By setting $s_{\ell+1} = k_{\mathtt{a}^\ell \mathtt{b}}$ the claim is proven. If $k_{\mathtt{a}^\ell \mathtt{b}} = \binom{k_{\mathtt{a}}}{\ell}(n - k_{\mathtt{a}})$, $s_{k_{\mathtt{a}}+1} = (n - k_{\mathtt{a}})$ and $s_i = 0$ for $i \in [k_{\mathtt{a}}]_0$ is the only possibility. Let secondly be $r \in [k_{\mathtt{b}}]_0$ and $k_{\mathtt{a}^\ell \mathtt{b}} = \binom{k_{\mathtt{a}}-1}{\ell} r + \binom{k_{\mathtt{a}}}{\ell}(n - k_{\mathtt{a}} - r)$ and suppose that $k_{\mathtt{a}^\ell \mathtt{b}}$ is not unique. This implies $s_{k_{\mathtt{a}}+1} < n - k_{\mathtt{a}} - r$. Assume that $s_{k_{\mathtt{a}}+1} = n - k_{\mathtt{a}} - r'$ for $r' \in [k_{\mathtt{b}}]_{>r}$. Thus there exists $x \in \mathbb{N}$ with $\binom{k_{\mathtt{a}}}{\ell}(n - k_{\mathtt{a}} - r') + x = \frac{(k_{\mathtt{a}}-1)!(k_{\mathtt{a}}(n-k_{\mathtt{a}})-\ell r)}{\ell!(k_{\mathtt{a}}-\ell)!}$, i.e., $x = \frac{(k_{\mathtt{a}}-1)!(k_{\mathtt{a}} r' - \ell r)}{\ell!(k_{\mathtt{a}}-\ell)!}$. By $k_{\mathtt{b}} = n - k_{\mathtt{a}}$ we have $x \leq \binom{k_{\mathtt{a}}-1}{\ell} r' = \frac{(k_{\mathtt{a}}-1)!(k_{\mathtt{a}} r' - \ell r')}{\ell!(k_{\mathtt{a}}-\ell)!}$ (we only have r' occurrences of $\mathtt{b}$ left to distribute). By $r' > r$ we have $\frac{(k_{\mathtt{a}}-1)!(k_{\mathtt{a}} r' - \ell r)}{\ell!(k_{\mathtt{a}}-\ell)!} = x < \frac{(k_{\mathtt{a}}-1)!(k_{\mathtt{a}} r' - \ell r)}{\ell!(k_{\mathtt{a}}-\ell)!}$ - a contradiction. □

Since we are not able to fully characterise the uniquely determined values for each $k_{\mathtt{a}^\ell \mathtt{b}}$ for arbitrary n and ℓ, the following proposition gives the characterisation for $\ell \in \{0, 1\}$. Notice that we use $k_{\mathtt{a}}$ immediately since it is determinable by n and $k_{\mathtt{a}^0 \mathtt{b}} = k_{\mathtt{b}}$.

Proposition 17 (∗). *The word* $w \in \Sigma^n$ *is uniquely determined by* $k_{\mathtt{a}}$ *and* $k_{\mathtt{ab}}$ *iff one of the following occurs*

- $k_{\mathtt{a}} = 0$ *or* $k_{\mathtt{a}} = n$ *(and obviously* $k_{\mathtt{ab}} = 0$*),*

– $k_{\mathtt{a}} = 1$ *or* $k_{\mathtt{a}} = n-1$ *and* $k_{\mathtt{ab}}$ *is arbitrary,*
– $k_{\mathtt{a}} \in [n-2]_{\geq 2}$ *and* $k_{\mathtt{ab}} \in \{0, 1, k_{\mathtt{a}}(n-k_{\mathtt{a}})-1, k_{\mathtt{a}}(n-k_{\mathtt{a}})\}$.

In all cases not covered by Proposition 17 the word cannot be uniquely determined by $\binom{w}{\mathtt{a}}$ and $\binom{w}{\mathtt{ab}}$. The following theorem combines the reconstruction of a word with the binomial coefficients of right-bounded-block words.

Theorem 18. *Let* $j \in [k_{\mathtt{a}}]_0$. *If* $k_{\mathtt{a}^j\mathtt{b}}$ *is unique, then the word* $w \in \Sigma^n$ *is uniquely determined by* $\{\mathtt{b}, \mathtt{ab}, \mathtt{a}^2\mathtt{b}, \ldots, \mathtt{a}^j\mathtt{b}\}$.

Proof. If $k_{\mathtt{a}^j\mathtt{b}}$ is unique, the coefficients $s_{j+1}, \ldots, s_{k_{\mathtt{a}}+1}$ are uniquely determined. Substituting backwards the known values in the first $j-1$ equations (2) (for $\ell = 1, \ldots, j-1$) we can now obtain successively the values for $s_j, \ldots, s_1$. □

Corollary 19. *Let* ℓ *be minimal such that* $k_{\mathtt{a}^\ell\mathtt{b}}$ *is unique. Then* w *is uniquely determined by* $\{\mathtt{a}, \mathtt{ab}, \mathtt{a}^2\mathtt{b}, \ldots, \mathtt{a}^\ell\mathtt{b}\}$ *and not uniquely determined by any* $\{\mathtt{a}, \mathtt{ab}, \mathtt{a}^2\mathtt{b}, \ldots, \mathtt{a}^i\mathtt{b}\}$ *for* $i < \ell$.

Proof. It follows directly from Theorem 18. □

By [15] an upper bound on the number of binomial coefficients to uniquely reconstruct the word $w \in \Sigma^n$ is given by the amount of the binomial coefficients of the $(\lfloor \frac{16}{7}\sqrt{n} \rfloor + 5)$-spectrum. Notice that implicitly the full spectrum is assumed to be known. As proven in Section 2, Lyndon words up to this length suffice. Since there are $\frac{1}{n}\sum_{d|n} \mu(d) \cdot 2^{\frac{n}{d}}$ Lyndon words of length n, the combination of both results presented in [15, 22] states that, for $n > 6$,

$$\sum_{i=1}^{\lfloor \frac{16}{7}\sqrt{n} \rfloor + 5} \frac{1}{i} \sum_{d|i} \mu(d) \cdot 2^{\frac{i}{d}} \tag{3}$$

binomial coefficients are sufficient for a unique reconstruction with the Möbius function μ. Up to now, it was the best known upper bound.

Theorem 18 shows that $\min\{k_{\mathtt{a}}, k_{\mathtt{b}}\} + 1$ binomial coefficients are enough for reconstructing a binary word uniquely. By Proposition 17 we need exactly one binomial coefficient if $n \in [3]$ and at most two if $n = 4$. For $n \in \{5, 6\}$ we need at most $n-2$ different binomial coefficients. The following theorem shows that by Theorem 18 we need strictly less binomial coefficients for $n > 6$.

Theorem 20 (∗). *Let* $w \in \Sigma^n$. *We have that* $\min\{k_{\mathtt{a}}, k_{\mathtt{b}}\} + 1$ *binomial coefficients suffice to uniquely reconstruct* w. *If* $k_{\mathtt{a}} \leq k_{\mathtt{b}}$, *then the set of sufficient binomial coefficients is* $S = \{\mathtt{b}, \mathtt{ab}, \mathtt{a}^2\mathtt{b}, ..., \mathtt{a}^h\mathtt{b}\}$ *where* $h = \lfloor \frac{n}{2} \rfloor$. *If* $k_{\mathtt{a}} > k_{\mathtt{b}}$, *then the set is* $S = \{\mathtt{a}, \mathtt{ba}, \mathtt{b}^2\mathtt{a}, ..., \mathtt{b}^h\mathtt{a}\}$. *This bound is strictly smaller than (3).*

Remark 21. By Lemma 16 we know that $k_{\mathtt{a}^\ell\mathtt{b}}$ is unique if it is in $[\ell]_0$ or exactly $\binom{k_{\mathtt{a}}}{\ell}(n-k_{\mathtt{a}})$. The probability for the latter is $\frac{1}{2^n}$ for $w \in \{\mathtt{a}, \mathtt{b}\}^n$. If $k_{\mathtt{a}^\ell\mathtt{b}} = m \in [\ell]_0$ we get by (2) immediately $s_{\ell+1} = m$ and $s_i = 0$ for $\ell + 2 \leq i \leq k_{\mathtt{a}} + 1$. Hence, the values for s_j for $j \in [\ell]$ are not determined. By $\sum_{i\in[\ell]} s_i = n - k_{\mathtt{a}} - m$ there are $d = \sum_{i\in[\ell]_0} \binom{\ell}{\ell-i}\binom{n-k_{\mathtt{a}}-m-1}{i-1}$ possibilities to fulfill the constraints, i.e., we have a probability of $\frac{d}{2^n}$ to have such a word.

4 Reconstruction for Arbitrary Alphabets

In this section we address the problem of reconstructing words over arbitrary alphabets from their scattered factors. We begin with a series of results of algorithmic nature. Let $\Sigma = \{\mathtt{a}_1, \ldots, \mathtt{a}_q\}$ be an alphabet equipped with the ordering $\mathtt{a}_i < \mathtt{a}_j$ for $1 \leq i < j \leq q \in \mathbb{N}$.

Definition 22. *Let $w_1, \ldots, w_k \in \Sigma^*$ for $k \in \mathbb{N}$, and $K = (k_\mathtt{a})_{\mathtt{a}\in\Sigma}$ a sequence of $|\Sigma|$ natural numbers. A* $K-$valid marking *of $w_1, \ldots, w_k$ is a mapping $\psi : [k] \times \mathbb{N} \to \mathbb{N}$ such that for all $j \in [k]$, $i, \ell \in [|w_j|]$, and $\mathtt{a} \in \Sigma$ there holds*

- *if $w_j[i] = \mathtt{a}$ then $\psi(j, i) \leq k_\mathtt{a}$,*
- *if $i < \ell \leq |w_j|$ and $w_j[i] = w_j[\ell] = \mathtt{a}$ then $\psi(j, i) < \psi(j, \ell)$.*

A K-valid marking *of $w_1, \ldots, w_k$ is represented as the string w_1^ψ, w_2^ψ, …, w_k^ψ, where $w_j^\psi[i] = (w_j[i])_{\psi(j,i)}$ for fresh letters $(w_j[i])_{\psi(j,i)}$.*

For instance, let $k = 2$, $\Sigma = \{\mathtt{a}, \mathtt{b}\}$, and $w_1 = \mathtt{aab}$, $w_2 = \mathtt{abb}$. Let $k_\mathtt{a} = 3, k_\mathtt{b} = 2$ define the sequence K. A K-valid marking of w_1, w_2 would be $w_1^\psi = (\mathtt{a})_1(\mathtt{a})_3(\mathtt{b})_1, w_2^\psi = (\mathtt{a})_2(\mathtt{b})_1(\mathtt{b})_2$ defining ψ implicitly by the indices. We used parentheses in the marking of the letters in order to avoid confusions.

We recall that a topological sorting of a directed graph $G = (V, E)$, with $V = \{v_1, \ldots, v_n\}$, is a linear ordering $v_{\sigma(1)} < v_{\sigma(2)} < \ldots < v_{\sigma(n)}$ of the nodes, defined by the permutation $\sigma : [n] \to [n]$, such that there exists no edge in E from $v_{\sigma(i)}$ to $v_{\sigma(j)}$ for any $i > j$ (i.e., if v_a comes after v_b in the linear ordering, for some $a = \sigma(i)$ and $b = \sigma(j)$, then we have $i > j$ and there should be no edge between v_a and v_b). It is a folklore result that any directed graph G has a topological sorting if and only if G is acyclic.

Definition 23. *Let $w_1, \ldots, w_k \in \Sigma^*$ for $k \in \mathbb{N}$, $K = (k_\mathtt{a})_{\mathtt{a}\in\Sigma}$ a sequence of $|\Sigma|$ natural numbers, and ψ a* $K-$valid marking *of $w_1, \ldots, w_k$. Let G_ψ be the graph that has $\sum_{\mathtt{a}\in\Sigma} k_\mathtt{a}$ nodes, labelled with the letters $(\mathtt{a})_1, \ldots, (\mathtt{a})_{k_\mathtt{a}}$, for all $\mathtt{a} \in \Sigma$, and the directed edges $((w_j[i])_{\psi(j,i)}, (w_j[i+1])_{\psi(j,i+1)})$, for all $j \in [k]$, $i \in [|w_j|]$, and $((\mathtt{a})_i, (\mathtt{a})_{i+1})$, for all occuring i and $\mathtt{a} \in \Sigma$. We say that there exists* a valid topological sorting *of the ψ-marked letters of the words $w_1, \ldots, w_k$ if there exists a topological sorting of the nodes of G_ψ, i.e., G_ψ is a directed acyclic graph.*

The graph associated with the K-valid marking of w_1, w_2 from above would have the five nodes $(\mathtt{a})_1, (\mathtt{a})_2, (\mathtt{a})_3, (\mathtt{b})_1, (\mathtt{b})_2$ and the six directed edges $((\mathtt{a})_1, (\mathtt{a})_3)$, $((\mathtt{a})_3, (\mathtt{b})_1)$, $((\mathtt{a})_2, (\mathtt{b})_1)$, $((\mathtt{b})_1, (\mathtt{b})_2)$, $((\mathtt{a})_1, (\mathtt{a})_2), ((\mathtt{a})_2, (\mathtt{a})_3)$ (where the direction of the edge is from the left node to the right node of the pair defining it). This graph has the topological sorting $(\mathtt{a})_1(\mathtt{a})_2(\mathtt{a})_3(\mathtt{b})_1(\mathtt{b})_2$.

Theorem 24 (∗). *For $w_1, \ldots, w_k \in \Sigma^*$ and a sequence $K = (k_\mathtt{a})_{\mathtt{a}\in\Sigma}$ of $|\Sigma|$ natural numbers, there exists a word w such that w_i is a scattered factor of w with $|w|_\mathtt{a} = k_\mathtt{a}$, for all $i \in [k]$ and all $\mathtt{a} \in \Sigma$, if and only if there exist a K-valid marking ψ of the words $w_1, \ldots, w_k$ and a valid topological sorting of the ψ-marked letters of the words $w_1, \ldots, w_k$.*

Next we show that in Theorem 24 uniqueness propagates in the $\Leftarrow$-direction.

Corollary 25. *Let $w_1, \ldots, w_k \in \Sigma^*$ and $K = (k_{\mathtt{a}})_{\mathtt{a} \in \Sigma}$ a sequence of $|\Sigma|$ natural numbers. If the following hold*

- *there exists a unique K-valid marking ψ of the words $w_1, \ldots, w_k$,*
- *in the unique K-valid marking ψ we have that for each $\mathtt{a} \in \Sigma$ and $\ell \in [k_{\mathtt{a}}]$ there exists $i \in [k]$ and $j \in [|w_i|]$ with $\psi(i,j) = \ell$, and*
- *there exists a unique valid topological sorting of the ψ-marked letters of the words $w_1, \ldots, w_k$*

then there exists a unique word w such that w_i is a scattered factor of w, for all $i \in [k]$ and $|w|_{\mathtt{a}} = k_{\mathtt{a}}$ for all $\mathtt{a} \in \Sigma$.

Proof. Let w be the word obtained by writing in order the letters of the unique valid topological sorting of the ψ-marked letters of the words $w_1, \ldots, w_k$ and removing their markings. It is clear that w' has w_i as a scattered factor, for all $i \in [k]$, and that $|w|_{\mathtt{a}} = k_{\mathtt{a}}$, for all $\mathtt{a} \in \Sigma$. The word w is uniquely defined (as there is no other K-valid marking nor valid topological sorting of the ψ-marked letters), and $|w|_{\mathtt{a}} = k_{\mathtt{a}}$, for all $\mathtt{a} \in \Sigma$. □

In order to state the second result, we need the projection $\pi_S(w)$ of a word $w \in \Sigma^*$ on $S \subseteq \Sigma$: $\pi_S(w)$ is obtained from w by removing all letters from $\Sigma \setminus S$.

Theorem 26. *Set $W = \{w_{\mathtt{a},\mathtt{b}} \mid \mathtt{a} < \mathtt{b} \in \Sigma\}$ such that*

- *$w_{\mathtt{a},\mathtt{b}} \in \{\mathtt{a}, \mathtt{b}\}^*$ for all $\mathtt{a}, \mathtt{b} \in \Sigma$,*
- *for all $w, w' \in W$ and all $\mathtt{a} \in \Sigma$, if $|w|_{\mathtt{a}} \cdot |w'|_{\mathtt{a}} > 0$, then $|w|_{\mathtt{a}} = |w'|_{\mathtt{a}}$.*

Then there exists at most one $w \in \Sigma^*$ such that $w_{\mathtt{a},\mathtt{b}}$ is $\pi_{\{\mathtt{a},\mathtt{b}\}}(w)$ for all $\mathtt{a}, \mathtt{b} \in \Sigma$.

Proof. Notice firstly $|W| = \frac{q(q-1)}{2}$. Let $k_{\mathtt{a}} = |w_{\mathtt{a},\mathtt{b}}|_{\mathtt{a}}$, for $\mathtt{a} < \mathtt{b} \in \Sigma$. These numbers are clearly well defined, by the second item in our hypothesis. Let $K = (k_{\mathtt{a}})_{\mathtt{a} \in \Sigma}$. It is immediate that there exists a unique K-valid marking ψ of the words $(w_{\mathtt{a},\mathtt{b}})_{\mathtt{a}<\mathtt{b} \in \Sigma}$. As each two marked letters $(\mathtt{a})_i$ and $(\mathtt{b})_j$ (i.e., each two nodes $(\mathtt{a})_i$ and $(\mathtt{b})_j$ of G_ψ) appear in the marked word $w^{\psi}_{\mathtt{a},\mathtt{b}}$, we know the order in which these two nodes should occur in a topological sorting of G_ψ. This means that, if G_ψ is acyclic, then it has a unique topological sorting. Our statement follows now from Corollary 25. □

Remark 27. Given the set $W = \{w_{\mathtt{a},\mathtt{b}} \mid \mathtt{a} < \mathtt{b} \in \Sigma\}$ as in the statement of Theorem 26, with $k_{\mathtt{a}} = |w_{\mathtt{a},\mathtt{b}}|_{\mathtt{a}}$, for $\mathtt{a} < \mathtt{b} \in \Sigma$, and $K = (k_{\mathtt{a}})_{\mathtt{a} \in \Sigma}$, we can produce the unique K-valid marking ψ of the words $(w_{\mathtt{a},\mathtt{b}})_{\mathtt{a}<\mathtt{b} \in \Sigma}$ in linear time $O(\sum_{\mathtt{a}<\mathtt{b} \in \Sigma} |w_{\mathtt{a},\mathtt{b}}|) = O((q-1)\sum_{\mathtt{a} \in \Sigma} k_{\mathtt{a}})$: just replace the i^{th} letter $\mathtt{a}$ of $w_{\mathtt{a},\mathtt{b}}$ by $(\mathtt{a})_i$, for all $\mathtt{a}$ and i. The graph G_ψ has $O((q-1)\sum k_{\mathtt{a}})$ edges and $O(\sum k_{\mathtt{a}})$ vertices and can be constructed in linear time $O((q-1)\sum k_{\mathtt{a}})$. Sorting G_ψ topologically takes $O((q-1)\sum k_{\mathtt{a}})$ time (see, e.g., the handbook [4]). As such, we conclude that reconstructing a word $w \in \Sigma^*$ from its projections over all two-letter-subsets of Σ can be done in linear time w.r.t. the total length of the respective projections.

Theorem 26 is in a sense optimal: in order to reconstruct a word over Σ uniquely, we need all its projections on two-letter-subsets of Σ. That is, it is always the case that for a strict subset U of $\{\{\mathtt{a},\mathtt{b}\} \mid \mathtt{a} < \mathtt{b} \in \Sigma\}$, with $|U| = \frac{q(q-1)}{2} - 1$, there exist two words $w' \neq w$ such that $\{\pi_p(w') \mid p \in U\} = \{\pi_p(w) \mid p \in U\}$. We can, in fact, show the following results:

Theorem 28. *Let $S_1, \ldots, S_k$ be subsets of Σ. The following hold:*

1. *If each pair $\{\mathtt{a},\mathtt{b}\} \subseteq \Sigma$ is included in at least one of the sets S_i, then we can reconstruct any word uniquely from its projections $\pi_{S_1}(\cdot), \ldots, \pi_{S_k}(\cdot)$.*
2. *If there exists a pair $\{\mathtt{a},\mathtt{b}\}$ that is not contained in any of the sets $S_1, \ldots, S_k$, then there exist two words w and w' such that $w \neq w'$ and $\pi_{S_1}(w) = \pi_{S_1}(w'), \ldots, \pi_{S_k}(w) = \pi_{S_k}(w')$.*

Proof. The first part is, once again, a consequence of Corollary 25. The second part can be shown by assuming that $\Sigma = \{\mathtt{a}_1, \ldots, \mathtt{a}_q\}$ and the pair $\{\mathtt{a}_1, \mathtt{a}_2\}$ is not contained in any of the sets $S_1, \ldots, S_k$. Then, for $w = \mathtt{a}_1\mathtt{a}_3\mathtt{a}_4 \ldots \mathtt{a}_q$ and $w' = \mathtt{a}_2\mathtt{a}_3\mathtt{a}_4 \ldots \mathtt{a}_q$, we have that $\pi_{S_1}(w) = \pi_{S_1}(w'), \ldots, \pi_{S_k}(w) = \pi_{S_k}(w')$. □

In this context, we can ask how efficiently can we decide if a word is uniquely reconstructible from the projections $\pi_{S_1}(\cdot), \ldots, \pi_{S_k}(\cdot)$ for $S_1, \ldots, S_k \subset \Sigma$.

Theorem 29 (∗). *Given the sets $S_1, \ldots, S_k \subset \Sigma$, we decide whether we can reconstruct any word uniquely from its projections $\pi_{S_1}(\cdot), \ldots, \pi_{S_k}(\cdot)$ in $O(q^2k)$ time. Moreover, under the* Strong Exponential Time Hypothesis *(see the survey [3] and the references therein), there is no $O(q^{2-d}k^c)$ algorithm for solving the above decision problem, for any $d, c > 0$.*

Coming now back to combinatorial results, we use the method developed in Section 3 to reconstruct a word over an arbitrary alphabet. We show that we need at most $\sum_{i\in[q]} |w|_i(q+1-i)$ different binomial coefficients to reconstruct w uniquely for the alphabet $\Sigma = \{1, \ldots, q\}$. In fact, following the results from the first part of this section, we apply this method on all combinations of two letters. Consider for an example that for $w \in \{\mathtt{a},\mathtt{b},\mathtt{n}\}^6$ the following binomial coefficients $\binom{w}{\mathtt{a}^0\mathtt{b}} = 1$, $\binom{w}{\mathtt{a}^0\mathtt{n}} = 2$, $\binom{w}{\mathtt{a}^1\mathtt{b}} = 0$, $\binom{w}{\mathtt{a}^1\mathtt{n}} = 3$, $\binom{w}{\mathtt{b}^1\mathtt{n}} = 2$, and $\binom{w}{\mathtt{a}^2\mathtt{n}} = 1$ are given. By $|w| = 6$, $|w|_\mathtt{b} = 1$, and $|w|_\mathtt{n} = 2$, we get $|w|_\mathtt{a} = 3$. Applying the method from Section 3 for $\{\mathtt{a},\mathtt{b}\}$, $\{\mathtt{a},\mathtt{n}\}$, and $\{\mathtt{b},\mathtt{n}\}$ we obtain the scattered factors $\mathtt{ba}^3$, $\mathtt{anana}$, and $\mathtt{bn}^2$. Combining all these three scattered factors gives us uniquely $\mathtt{banana}$. Notice that in this example we only needed six binomial coefficients instead of ten, which is the worst case.

Remark 30. As seen in the example we have not only the word length but also $\binom{w}{\mathtt{x}}$ for all $\mathtt{x} \in \Sigma$ but one. Both information give us the remaining single letter binomial coefficient and hence we will assume that we know all of them.

For convenience in the following theorem consider $\Sigma = \{1, \ldots, q\}$ for $q > 2$ and set $\alpha := \lfloor \frac{16}{7}\sqrt{n} \rfloor + 5$. In the general case the results by [22] and [15] yield

that

$$\sum_{i\in[\alpha]} \frac{1}{i} \frac{(q+1)^{\frac{i}{2}}-1}{q} \tag{4}$$

is smaller than the best known upper bound on the number of binomial coefficients sufficient to reconstruct a word uniquely.

The following theorem generalises Theorem 20 on an arbitrary alphabet.

Theorem 31 (∗). *For reconstructing a word $w \in \Sigma^*$ of length at least $q-1$ uniquely, $\sum_{i\in[q]} |w|_i(q+1-i)$ binomial coefficients suffice, which is strictly smaller than (4).*

Remark 32. Since the estimation in Theorem 31 depends on the distribution of the letters in contrast to the method of reconstruction, it is wise to choose an order $<$ on Σ such that $x < y$ if $|w|_x \leq |w|_y$. In the example we have chosen the *natural* order $\mathtt{a} < \mathtt{b} < \mathtt{n}$ which leads in the worst case to fourteen binomial coefficients that has to be taken into consideration. If we chose the order $\mathtt{b} < \mathtt{n} < \mathtt{a}$ the formula from Theorem 31 provides that ten binomial coefficients suffice. This observation leads also to the fact that less binomial coefficients suffice for a unique determinism if the letters are not distributed equally but some letters occur very often and some only a few times.

Remark 33. Let's note that the number of binomial coefficients we need is at most qn. Indeed, we will prove that $\sum_{i\in[q]} |w|_i(q+1-i) \leq qn$. We have $qn = qn+n-n = q\sum_{i\in[q]} |w|_i + \sum_{i\in[q]} |w|_i - \sum_{i\in[q]} |w|_i \geq q\sum_{i\in[q]} |w|_i + \sum_{i\in[q]} |w|_i - \sum_{i\in[q]}(|w|_i i) = \sum_{i\in[q]} |w|_i(q+1-i)$.

5 Conclusion

In this paper we have proven that a relaxation of the so far investigated reconstruction problem from scattered factors from k-spectra to arbitrary sets yields that less scattered factors than the best known upper bound are sufficient to reconstruct a word uniquely. Not only in the binary but also in the general case the distribution of the letters plays an important role: in the binary case the amount of necessary binomial coefficients is smaller the larger $|w|_{\mathtt{a}} - |w|_{\mathtt{b}}$ is. The same observation results from the general case - if all letters are equally distributed in w then we need more binomial coefficients than in the case where some letters rarely occur and others occur much more often. Nevertheless the restriction to right-bounded-block words (that are intrinsically Lyndon words) shows that a word can be reconstructed by fewer binomial coefficients if scattered factors from different spectra are taken. Further investigations may lead into two directions: firstly a better characterisation of the uniqueness of the $k_{\mathtt{a}^\ell\mathtt{b}}$ would be helpful to understand better in which cases less than the worst case amount of binomial coefficients suffices and secondly other sets than the right-bounded-block words could be investigated for the reconstruction problem.

References

1. J. Berstel, J. Karhumäki, Combinatorics on Words – A Tutorial, *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS* **79** (2003), 178–228.
2. V. Berthé, J. Karhumäki, D. Nowotka, J. Shallit, Mini-Workshop: Combinatorics on Words, Oberwolfach Rep. **7** (2010), 2195–2244. doi: 10.4171/OWR/2010/37.
3. K. Bringmann: Fine-Grained Complexity Theory (Tutorial). 36th International Symposium on Theoretical Aspects of Computer Science (STACS 2019), R. Niedermeier, C. Paul Eds., *Leibniz International Proceedings in Informatics* **4** (2019), 1–7.
4. T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, *Introduction to Algorithms*, 3rd Edition. MIT Press (2009).
5. A. W. M. Dress, P. L. Erdős, Reconstructing words from subwords in linear time, *Ann. Comb.* **8** (2004), 457–462.
6. M. Dudik, L. J. Schulman, Reconstruction from subsequences, *J. Combin. Theory, Ser. A* **103** (2003), 337–348.
7. P. L. Erdős, P. Ligeti, P. Sziklai, D. C. Torney, Subwords in reverse-complement order, *Ann. Comb.* **10** (2006), 415–430.
8. M. Ferov, Irreducible Polynomial modulo p, Bachelor Thesis at Charles University Prague (2008).
9. P. Fleischmann, M. Lejeune, F. Manea, D. Nowotka, M. Rigo, Reconstructing words from right-bounded-block words (2020), arXiv:2001.11218, 21 pages.
10. D. D. Freydenberger, P. Gawrychowski, J. Karhumäki, F. Manea, W. Rytter, Testing k-binomial equivalence. Multidisciplinary Creativity: homage to G. Păun on his 65th birthday, 239–248, Ed. Spandugino, Bucharest, Romania (2015). Available as: arXiv:1509.00622.
11. F. Harary, On the reconstruction of a graph from a collection of subgraphs. In Theory of Graphs and its Applications (Proc. Sympos. Smolenice, 1963). Publ. House Czechoslovak Acad. Sci., Prague (1964), 47–52.
12. L. van Iersel, V. Moulton, Leaf-reconstructibility of phylogenetic networks, *SIAM J. Discrete Math.* **32** (2018), 2047–2066.
13. L. I. Kalashnik, The reconstruction of a word from fragments, in "Numerical Mathematics and Computer Technology", 56–57, Akad. Nauk Ukrain. SSR Inst. Mat., Preprint IV, (1973).
14. P. J. Kelly, A congruence theorem for trees, *Pacific J. Math.* **7** (1957), 961–968.
15. I. Krasikov, Y. Roditty, On a Reconstruction Problem for Sequences, *J. Combin. Theory, Ser. A* **77** (1997), 344–348.
16. V.I. Levenshtein, On perfect codes in deletion and insertion metric, *Discrete Math. Appl.* **2** (1992), 241–258.
17. M. Lothaire, *Combinatorics on Words*, Cambridge University Press (1997).
18. J. Manŭch, Characterization of a Word by its Subwords, in Developments in Language Theory (1999), 210–219, World Scientific.
19. B. Manvel, A. Meyerowitz, A. Schwenk, K. Smith, P. Stockmeyer, Reconstruction of sequences, *Discrete Math.* **94** (1991), 209–219.
20. B. Manvel, P. K. Stockmeyer, On reconstruction of matrices, *Math. Mag.* **44** (1971), 218–221.
21. P. V. O'Neil, Ulam's conjecture and graph reconstructions, *Amer. Math. Monthly* **77** (1970), 35–43.
22. C. Reutenauer, *Free Lie Algebras*, London Mathematical Society Monographs New Series, P. M. Cohn, H. G. Dales Eds., (1993).

23. M. Rigo, P. Salimov, Another generalization of abelian equivalence: Binomial complexity of infinite words, *Theoret. Comp. Sci.* **601** (2015), 47–57.
24. G. Rozenberg, A. Salomaa, *Handbook of Formal Languages* (3 volumes), Springer (1997).
25. I. Simon, Piecewise testable events, In: Automata Theory and Formal Languages, H. Brakhage ed., *Lect. Notes Comp. Sci.* **33** (1975), 214–222.

Appendix for Section 2

Example for Binomial Coefficients computed by Lyndon Words.

Let us recall that, for computing $\binom{w}{u}$, x and y are words such that $u = xy$ and such that every word appearing in the polynomial $x \sqcup\!\sqcup y$ is lexicographically less than or equal to u.

Example 34. Considering $\Sigma = \{\mathtt{a}, \mathtt{b}\}$ the binomial coefficient $\binom{w}{\mathtt{ba}}$ can be computed using the Lyndon words $\mathtt{a}$, and $\mathtt{b}$ by

$$\binom{w}{\mathtt{ba}} = \frac{1}{(\mathtt{b} \sqcup\!\sqcup \mathtt{a}, \mathtt{ba})}\left[\binom{w}{\mathtt{b}}\binom{w}{\mathtt{a}} - (\mathtt{b}\downarrow \mathtt{a}, \mathtt{ab})\binom{w}{\mathtt{ab}}\right] = \binom{w}{\mathtt{b}}\binom{w}{\mathtt{a}} - \binom{w}{\mathtt{ab}}.$$

Regarding word length three, the Lyndon words are $\mathtt{aab}$ and $\mathtt{abb}$. Let us give formulas to compute $\binom{w}{\mathtt{aba}}$, $\binom{w}{\mathtt{baa}}$, $\binom{w}{\mathtt{bab}}$ and $\binom{w}{\mathtt{bba}}$. Having $x = \mathtt{ab}$ and $y = \mathtt{a}$, we obtain

$$\binom{w}{\mathtt{aba}} = \binom{w}{\mathtt{ab}}\left[\binom{w}{\mathtt{a}} - 1\right] - 2\binom{w}{\mathtt{aab}}.$$

For $u = \mathtt{baa}$, we can either choose $x = \mathtt{b}$ and $y = \mathtt{aa}$ or $x = \mathtt{ba}$ and $y = \mathtt{a}$. In the first case, we get

$$\binom{w}{\mathtt{baa}} = \binom{w}{\mathtt{b}}\binom{w}{\mathtt{aa}} - \binom{w}{\mathtt{aba}} - \binom{w}{\mathtt{aab}}$$

and by reinjecting formulas for $\binom{w}{\mathtt{aa}}$ and $\binom{w}{\mathtt{aba}}$, obtained recursively,

$$\binom{w}{\mathtt{baa}} = \left[\binom{w}{\mathtt{a}} - 1\right]\left[\frac{1}{2}\binom{w}{\mathtt{a}}\binom{w}{\mathtt{b}} - \binom{w}{\mathtt{ab}}\right] + \binom{w}{\mathtt{aab}}.$$

Finally, the last two formulas are quite similar to what we already had:

$$\binom{w}{\mathtt{bab}} = \binom{w}{\mathtt{ab}}\left[\binom{w}{\mathtt{b}} - 1\right] - 2\binom{w}{\mathtt{abb}}$$

and

$$\binom{w}{\mathtt{bba}} = \left[\binom{w}{\mathtt{b}} - 1\right]\left[\frac{1}{2}\binom{w}{\mathtt{a}}\binom{w}{\mathtt{b}} - \binom{w}{\mathtt{ab}}\right] + \binom{w}{\mathtt{abb}}.$$

Appendix for Section 3

Proposition 17. *The word $w \in \Sigma^n$ is uniquely determined by $k_{\mathtt{a}}$ and $k_{\mathtt{ab}}$ iff one of the following occurs*

- $k_{\mathtt{a}} = 0$ *or* $k_{\mathtt{a}} = n$ *(and obviously* $k_{\mathtt{ab}} = 0$*),*
- $k_{\mathtt{a}} = 1$ *or* $k_{\mathtt{a}} = n - 1$ *and* $k_{\mathtt{ab}}$ *is arbitrary,*
- $k_{\mathtt{a}} \in [n-2]_{\geq 2}$ *and* $k_{\mathtt{ab}} \in \{0, 1, k_{\mathtt{a}}(n - k_{\mathtt{a}}) - 1, k_{\mathtt{a}}(n - k_{\mathtt{a}})\}$.

Proof. Let us first prove that w is uniquely determined in these cases. It is obvious if $k_{\mathtt{a}} = 0$ or $k_{\mathtt{a}} = n$ since the word is composed of the same letter repeated n times. If $k_{\mathtt{a}} = 1$, then $w = \mathtt{b}^{s_1}\mathtt{ab}^{n-1-s_1}$ and $\binom{w}{\mathtt{ab}} = n-1-s_1 = k_{\mathtt{ab}}$. Therefore w is uniquely determined. If $k_{\mathtt{a}} = n-1$, then $w = \mathtt{b}^{s_1}\mathtt{ab}^{s_2}\cdots\mathtt{ab}^{s_n}$ with exactly one of the s_i being non zero and, in fact, equal to one. We have $\binom{w}{\mathtt{ab}} = \sum_{i=2}^{n}(i-1)s_i$ and, if $k_{\mathtt{ab}}$ is given (between 0 and $n-1$), then $s_{k_{\mathtt{ab}}+1} = 1$ is the only non zero exponent. Consider now $k_{\mathtt{a}} \in [n-2]_{\geq 2}$, i.e. $w = \mathtt{b}^{s_1}\mathtt{ab}^{s_2}\ldots\mathtt{b}^{s_{k_{\mathtt{a}}}}\mathtt{ab}^{s_{k_{\mathtt{a}}+1}}$. Thus $k_{\mathtt{ab}} = 0$ implies $s_1 = n - k_{\mathtt{a}}$ and $s_2 = 0, \ldots, s_{k_{\mathtt{a}}+1} = 0$ while $k_{\mathtt{ab}} = 1$ implies $s_2 = 1$, $s_1 = n-k_{\mathtt{a}}-1$ and $s_3 = 0, \ldots, s_{k_{\mathtt{a}}+1} = 0$. By Lemma 14, we know that (2) is maximal if and only if $s_{k_{\mathtt{a}}+1} = n - k_{\mathtt{a}}$ and all the other s_i are equal to zero. In that case, the value of the sum equals $k_{\mathtt{a}}(n-k_{\mathtt{a}})$. Therefore, if $\binom{w}{\mathtt{ab}} = k_{\mathtt{a}}(n-k_{\mathtt{a}})$, the word w is uniquely determined. Finally, if $k_{\mathtt{ab}} = k_{\mathtt{a}}(n-k_{\mathtt{a}}) - 1$, we must have $s_{k_{\mathtt{a}}+1} \leq n - k_{\mathtt{a}} - 1$. If we choose $s_{k_{\mathtt{a}}+1} = n - k_{\mathtt{a}} - 1$, it remains that $\sum_{i=1}^{k_{\mathtt{a}}} s_i = 1$ and $\sum_{i=2}^{k_{\mathtt{a}}}(i-1)s_i = k_{\mathtt{a}} - 1$. We must have $s_{k_{\mathtt{a}}} = 1$ and the other ones equal to zero. In fact, choosing $s_{k_{\mathtt{a}}+1} = n - k_{\mathtt{a}} - 1$ is the only possibility: if otherwise $s_{k_{\mathtt{a}}+1} = n - k_{\mathtt{a}} - \ell$ with $\ell > 1$, we obtain that $\sum_{i=2}^{k_{\mathtt{a}}}(i-1)s_i \geq \ell k_{\mathtt{a}} - 1$ with $\sum_{i=1}^{k_{\mathtt{a}}} s_i = \ell$. It is easy to check with Lemma 14 that these conditions are incompatible.

We now need to prove that w cannot be uniquely determined if $k_{\mathtt{a}} \in [n-2]_{\geq 2}$ and $k_{\mathtt{ab}} \in [k_{\mathtt{a}}(n-k_{\mathtt{a}})-2]_{\geq 2}$. To this aim we will give two different sets of values for the s_i. The first decomposition is the greedy one. Let us put $s_{k_{\mathtt{a}}+1} = \lfloor \frac{k_{\mathtt{ab}}}{k_{\mathtt{a}}} \rfloor$, $s_{(k_{\mathtt{ab}} \bmod k_{\mathtt{a}})+1} = 1$ and the other s_i equal to 0. Let us finally modify the value of s_1 (which is, at this stage, equal to 0 or 1) by adding the value needed. By $\sum_{i=1}^{k_{\mathtt{a}}+1} s_i = n - k_{\mathtt{a}}$ we get $s_1 \leftarrow s_1 + (n-k_{\mathtt{a}}) - \lfloor \frac{k_{\mathtt{ab}}}{k_{\mathtt{a}}} \rfloor - 1$. This implies $\sum_{i=1}^{k_{\mathtt{a}}+1} s_i = 1 + (n-k_{\mathtt{a}}) - \left\lfloor \frac{k_{\mathtt{ab}}}{k_{\mathtt{a}}} \right\rfloor - 1 + \left\lfloor \frac{k_{\mathtt{ab}}}{k_{\mathtt{a}}} \right\rfloor = n - k_{\mathtt{a}}$ and $s_i \geq 0$ for all i. Moreover we have $\sum_{i=2}^{k_{\mathtt{a}}+1}(i-1)s_i = (k_{\mathtt{ab}} \bmod k_{\mathtt{a}}) + k_{\mathtt{a}} \left\lfloor \frac{k_{\mathtt{ab}}}{k_{\mathtt{a}}} \right\rfloor = k_{\mathtt{ab}}$.

Now we provide a second decomposition for the s_i. First, let us assume that $2 \leq k_{\mathtt{ab}} < k_{\mathtt{a}}$. In that case, the greedy algorithm sets $s_{k_{\mathtt{ab}}+1} = 1$, $s_1 = n - k_{\mathtt{a}} - 1$ and the other s_i to 0. Let us now set $s_1 = n - k_{\mathtt{a}} - 2$ and all the other s_i to 0. Then, update $s_{k_{\mathtt{ab}}} \leftarrow s_{k_{\mathtt{ab}}} + 1$ and $s_2 \leftarrow s_2 + 1$ (in the case where $k_{\mathtt{ab}} = 2$, s_2 will be equal to 2 after these manipulations). We have that the sum in (2) is equal to $1 + (k_{\mathtt{ab}} - 1)$ as needed. Finally, if $k_{\mathtt{ab}} \geq k_{\mathtt{a}}$, then $s_{k_{\mathtt{a}}+1}$ was non zero in the greedy decomposition, and the idea is to reduce it of a value 1. Let us set $s_{k_{\mathtt{a}}+1} = \lfloor \frac{k_{\mathtt{ab}}}{k_{\mathtt{a}}} \rfloor - 1$ and the other s_i to 0. Then, let us update some values: $s_{(k_{\mathtt{ab}} \bmod k_{\mathtt{a}})+2} \leftarrow s_{(k_{\mathtt{ab}} \bmod k_{\mathtt{a}})+2} + 1$ and $s_{k_{\mathtt{a}}} \leftarrow s_{k_{\mathtt{a}}} + 1$ if $(k_{\mathtt{ab}} \bmod k_{\mathtt{a}}) \neq k_{\mathtt{a}} - 1$, and $s_{k_{\mathtt{a}}} = 2$, $s_2 = 1$ otherwise. Finally, set s_1 to the right value, i.e., $n - k_{\mathtt{a}} - \sum_{i=2}^{k_{\mathtt{a}}+1} s_i$. It can be easily checked that, in both cases, $s_1 \geq 0$ (notice that $(k_{\mathtt{ab}} \bmod k_{\mathtt{a}}) = k_{\mathtt{a}} - 1$ implies that $\lfloor \frac{k_{\mathtt{ab}}}{k_{\mathtt{a}}} \rfloor \leq n - k_{\mathtt{a}} - 2$) and that all s_i sum up to $n - k_{\mathtt{a}}$. Similarly, we can check that $\sum_{i=2}^{k_{\mathtt{a}}+1}(i-1)s_i$ is equal to $k_{\mathtt{ab}}$ in both cases.

To sum up, we gave two different decompositions for the s_i in cases where $k_{\mathtt{a}} \in [n-2]_{\geq 2}$ and $k_{\mathtt{ab}} \in [k_{\mathtt{a}}(n-k_{\mathtt{a}})-2]_{\geq 2}$. That implies that w cannot be uniquely determined in those cases. □

Theorem 20. *Let $w \in \Sigma^n$. We have that $\min\{k_{\mathtt{a}}, k_{\mathtt{b}}\} + 1$ binomial coefficients suffice to uniquely reconstruct w. If $k_{\mathtt{a}} \leq k_{\mathtt{b}}$, then the set of sufficient binomial coefficients is $S = \{\mathtt{b}, \mathtt{ab}, \mathtt{a}^2\mathtt{b}, ..., \mathtt{a}^h\mathtt{b}\}$ where $h = \lfloor \frac{n}{2} \rfloor$. If $k_{\mathtt{a}} > k_{\mathtt{b}}$, then the set is $S = \{\mathtt{a}, \mathtt{ba}, \mathtt{b}^2\mathtt{a}, ..., \mathtt{b}^h\mathtt{a}\}$. This bound is strictly smaller than (3).*

Proof. Assume w.l.o.g. $k_{\mathtt{a}} \leq k_{\mathtt{b}}$. Then $k_{\mathtt{a}} \leq \frac{n}{2}$ and Theorem 18 shows that words in the set $\{\mathtt{b}, \mathtt{ab}, \ldots, \mathtt{a}^{\lfloor \frac{n}{2} \rfloor}\mathtt{b}\}$ can reconstruct w uniquely. If $k_{\mathtt{a}} > k_{\mathtt{b}}$, the set S is obtained by replacing the letter $\mathtt{a}$ by $\mathtt{b}$ and vice-versa.

Set $N_2(i) := \frac{1}{i} \sum_{d|i} \mu(d) 2^{\frac{i}{d}}$ for all $i \in [\lfloor \frac{16}{7}\sqrt{n} \rfloor + 5]$, i.e., $\sum_{i=1}^{\lfloor \frac{16}{7}\sqrt{n} \rfloor + 5} N_2(i)$, which is Equation (3), binomial coefficients suffice. By [8, Lemma 2.4] we have

$$N_2(i) \geq \frac{1}{i}\left(2^i - \frac{2^{\frac{i}{2}} - 1}{2 - 1}\right) = \frac{1}{i}\left(2^i - 2^{\frac{i}{2}} + 1\right) = \frac{1}{i}\left(2^{\frac{i}{2}}(2^{\frac{i}{2}} - 1) + 1\right) \geq \frac{2^{\frac{i}{2}}}{i}.$$

This results in

$$\sum_{i=1}^{\lfloor \frac{16}{7}\sqrt{n} \rfloor + 5} N_2(i) \geq \sum_{i=1}^{\lfloor \frac{16}{7}\sqrt{n} \rfloor + 5} \frac{2^{\frac{i}{2}}}{i} \geq \frac{1}{\frac{16}{7}\sqrt{n} + 5} \frac{\sqrt{2}^{\frac{16}{7}\sqrt{n}+5} - 1}{\sqrt{2} - 1}.$$

We want to show that this quantity is at least equal to $\frac{n+1}{2}$. Let us define

$$f(x) = \frac{1}{\frac{16}{7}\sqrt{x} + 5} \frac{\sqrt{2}^{\frac{16}{7}\sqrt{x}+5} - 1}{\sqrt{2} - 1} - \frac{x+1}{2}$$

for all $x > 0$, which is the continuous extension on $\mathbb{R}^+$ of the quantity we are interested in. It is easy to verify by hand that $f(1), f(2), f(3)$ and $f(4)$ are positive. Let us formally show that $f(x) > 0$ for all $x \geq 5$. Since this function is differentiable, we get with $y = \frac{16}{7}\sqrt{x} + 5$

$$f'(x) = \frac{1}{y}\sqrt{2}^y \frac{\ln(\sqrt{2})}{\sqrt{2} - 1} \frac{8}{7\sqrt{x}} - \frac{1}{y^2} \frac{8}{7\sqrt{x}} \frac{\sqrt{2}^y - 1}{\sqrt{2} - 1} - \frac{1}{2}.$$

We thus have $\sqrt{x} = \frac{7y-35}{16}$ and $y \geq 7$ for all $x \geq 1$. By injecting y in the previous expression, and reducing to the common denominator, we have to show that

$$\begin{aligned}
&2y\sqrt{2}^y 128 \ln(\sqrt{2}) - 256(\sqrt{2}^y - 1) - 7(7y - 35)y^2(\sqrt{2} - 1) \\
&= \sqrt{2}^y(128 \ln(2)y - 256) + 256 - 49y^3(\sqrt{2} - 1) + 245y^2(\sqrt{2} - 1) \\
&\geq \sqrt{2}^y 365 + 256 - 49y^3(\sqrt{2} - 1) + 245y^2(\sqrt{2} - 1)
\end{aligned}$$

is strictly positive. Let us call the last quantity $g(y)$. We will show that it is positive for all $y \geq 10.05$, which means that $f(x)$ is positive for all x such that $\frac{16}{7}\sqrt{x} + 5 \geq 10.05$, i.e., for all $x \geq 5$. We have

$$\begin{aligned}
g'(y) &= 365\sqrt{2}^y \ln(\sqrt{2}) - 147(\sqrt{2} - 1)y^2 + 490(\sqrt{2} - 1)y, \\
g''(y) &= 365\sqrt{2}^y (\ln(\sqrt{2}))^2 - 294(\sqrt{2} - 1)y + 490(\sqrt{2} - 1), \\
g'''(y) &= 365\sqrt{2}^y (\ln(\sqrt{2}))^3 - 294(\sqrt{2} - 1),
\end{aligned}$$

and $g'''(7) > 50$, $g''(8.5) > 2$, $g'(10.05) > 8$ and finally $g(10.05) > 1787$. Since $g'''(y)$ is increasing and positive in 7, $g''(y)$ is increasing for $y \geq 7$. Therefore $g'(y)$ is increasing for $y \geq 8.5$ and finally $g(y)$ is increasing for $y \geq 10.05$ and positive. □

Appendix for Section 4

Theorem 24. *For $w_1, \ldots, w_k \in \Sigma^*$ and a sequence $K = (k_{\mathtt{a}})_{\mathtt{a} \in \Sigma}$ of $|\Sigma|$ natural numbers, there exists a word w such that w_i is a scattered factor of w with $|w|_{\mathtt{a}} = k_{\mathtt{a}}$, for all $i \in [k]$ and all $\mathtt{a} \in \Sigma$, if and only if there exist a K-valid marking ψ of the words $w_1, \ldots, w_k$ and a valid topological sorting of the ψ-marked letters of the words $w_1, \ldots, w_k$.*

Proof. If w is such that w_i is a scattered factor of w, for all $i \in [k]$, and $|w|_{\mathtt{a}} = k_{\mathtt{a}}$, for all $\mathtt{a} \in \Sigma$, then we can mark the i^{th} occurrence of $\mathtt{a}$ as $(\mathtt{a})_i$, for all $\mathtt{a} \in \Sigma$ and $i \in [k_{\mathtt{a}}]$. This induces a K-valid marking ψ of the words w_i, and, moreover, the linear ordering of the nodes of G_ψ induced by the order in which the marked letters (i.e., nodes of G_ψ) occur in w is a topological sorting of G_ψ.

Let us now assume that there exists a K-valid marking ψ of the words $w_1, \ldots, w_k$, and there exists a valid topological sorting of the ψ-marked letters of the words $w_1, \ldots, w_k$. Let w' be the word obtained by writing the nodes of G_ψ in the order given by its topological sorting and removing their markings. It is clear that w' has w_i as a scattered factor, for all $i \in [k]$, and that $|w'|_{\mathtt{a}} \leq k_{\mathtt{a}}$, for all $\mathtt{a} \in \Sigma$. Let now $w = w' \prod_{\mathtt{a} \in \Sigma} \mathtt{a}^{k_{\mathtt{a}} - |w'|_{\mathtt{a}}}$, where $\prod_{\mathtt{a} \in \Sigma} \mathtt{a}^{k_{\mathtt{a}} - |w'|_{\mathtt{a}}}$ is the concatenation of the factors $\mathtt{a}^{k_{\mathtt{a}} - |w'|_{\mathtt{a}}}$, for $\mathtt{a} \in \Sigma$ in some fixed order. Now w has w_i as a scattered factor, for all $i \in [k]$, and $|w|_{\mathtt{a}} = k_{\mathtt{a}}$, for all $\mathtt{a} \in \Sigma$. □

Theorem 29. *Given the sets $S_1, \ldots, S_k \subset \Sigma$, we decide whether we can reconstruct any word uniquely from its projections $\pi_{S_1}(\cdot), \ldots, \pi_{S_k}(\cdot)$ in $O(q^2 k)$ time. Moreover, under the* Strong Exponential Time Hypothesis *(see the survey [3] and the references therein), there is no $O(q^{2-d} k^c)$ algorithm for solving the above decision problem, for any $d, c > 0$.*

Proof. We begin with a series of preliminaries. Let us recall the *Orthogonal Vectors* problem: Given sets A, B consisting of n vectors in $\{0,1\}^k$, decide whether there are vectors $a \in A$ and $b \in B$ which are orthogonal (i.e., for any $i \in [k]$ we have $a[i]b[i] = 0$). This problem can be solved naïvely in $O(n^2 k)$ time, but under the *Strong Exponential Time Hypothesis* there is no $O(n^{2-d} k^c)$ algorithm for solving it, for any $d, c > 0$ (once more, see the survey [3] and the references therein).

We show that our problem is equivalent to the *Orthogonal Vectors* problem.

Let us first assume that we are given the sets $S_1, \ldots, S_k \subset \Sigma$, and we want to decide whether we can reconstruct any word uniquely from its projections $\pi_{S_1}(\cdot), \ldots, \pi_{S_k}(\cdot)$. This is equivalent, according to Theorem 28, to checking whether each pair $\{\mathtt{a}, \mathtt{b}\} \subseteq \Sigma$ is included in at least one of the sets S_i. For each letter $\mathtt{a}$ of Σ we define the k-dimensional vectors $x_{\mathtt{a}}$ where $x_{\mathtt{a}}[i] = 1$ if $\mathtt{a} \in S_i$ and

$x_{\mathtt{a}}[i] = 0$ if $\mathtt{a} \notin S_i$. This can be clearly done in $O(qk)$ time. Now, there exists a pair $\{\mathtt{a}, \mathtt{b}\}$ that is not contained in any of the sets $S_1, \ldots, S_k$ if and only if there exists a pair of vectors $\{x_{\mathtt{a}}, x_{\mathtt{b}}\}$ such that $x_{\mathtt{a}}[i]x_{\mathtt{b}}[i] = 0$ for all $i \in [k]$. We can check whether there exists a pair of vectors $\{x_{\mathtt{a}}, x_{\mathtt{b}}\}$ such that $x_{\mathtt{a}}[i]x_{\mathtt{b}}[i] = 0$ for all $i \in [k]$ by solving the *Orthogonal Vectors* problem by using for both input sets of vectors the set $\{x_{\mathtt{a}} \mid \mathtt{a} \in \Sigma\}$. As such, we can check whether there exists a pair $\{\mathtt{a}, \mathtt{b}\}$ that is not contained in any of the sets $S_1, \ldots, S_k$ in $O(q^2k)$ time.

Let us now assume that we are given two sets A, B consisting of n vectors in $\{0,1\}^k$, and we want to decide whether there are vectors $a \in A$ and $b \in B$ which are orthogonal (i.e., for any $i \in [k]$ we have $a[i]b[i] = 0$). We can compute the set of $(k+2)$-dimensional vectors A' containing the vectors of A extended with two new positions (position $k+1$ and position $k+2$) set to 10 and the vectors of B extended with two new positions (position $k+1$ and position $k+2$) set to 01. To decide whether there are vectors $a \in A$ and $b \in B$ which are orthogonal is equivalent to decide whether there are vectors $a, b \in A'$ which are orthogonal (if two such vectors exist, they must be different on their last two positions, so one must come from A and one from B). Assume that $A' = \{x_1, x_2, \ldots, x_{2n}\}$. Now we define an alphabet $\Sigma = \{\mathtt{a}_1, \ldots, \mathtt{a}_{2n}\}$ of size $2n$ and the sets $S_1, \ldots, S_k$, where $\mathtt{a}_j \in S_i$ if and only if $x_j[i] = 1$. Computing A' and then the alphabet Σ and the sets S_i, for $i \in [k]$, takes $O(nk)$ time. Now, to decide whether there are vectors $x_i, x_j \in A'$ which are orthogonal is equivalent to decide whether there exists a pair of letters $\{\mathtt{a}_i, \mathtt{a}_j\}$ of Σ that is not contained in any of the sets $S_1, \ldots, S_k$. The conclusion of the theorem now follows. □

Theorem 31. *For reconstructing a word $w \in \Sigma^*$ of length at least $q-1$ uniquely, $\sum_{i\in[q]} |w|_i(q+1-i)$ binomial coefficients suffice, which is strictly smaller than (4).*

Proof. The claim that $\sum_{i\in[q]} |w|_i(q+1-i)$ binomial coefficients suffice to reconstruct w uniquely follows by Theorem 26: for each pair of letters we apply the method of the binary case. We are thus going to reconstruct words $w_{\mathtt{a},\mathtt{b}}$ for all pairs of letters $\mathtt{a} < \mathtt{b}$. To determine such a word uniquely, $\min(k_{\mathtt{a}}, k_{\mathtt{b}}) + 1 \leq k_{\mathtt{a}} + 1$ binomial coefficients suffice. In total, we thus need $\sum_{i\in[q]}(|w|_i + 1)(q-i)$ binomial coefficients. Among them are the ones related to letters of $\Sigma\backslash\{q\}$. The letter i is counted $(q-i)$ times in the above formula. So in fact $\sum_{i\in[q]} |w|_i(q-i) + (q-1)$ binomial coefficients suffice. This quantity is less than or equal to $\sum_{i\in[q]} |w|_i(q+1-i)$ for every w of length at least $q-1$.

We show the second claim by induction on q where the binary case in Theorem 20 serves as induction basis. This implies

$$\begin{aligned}\sum_{i\in[q]} |w|_i(q+1-i) &= \left(\sum_{i\in[q-1]} |w|_i(q+1-i)\right) + |w|_q(q+1-q) \\ &= \left(\sum_{i\in[q-1]} |w|_i(q-i)\right) + \sum_{i\in[q-1]} |w|_i + |w|_q\end{aligned}$$

and therefore

$$\sum_{i\in[q]} |w|_i(q+1-i) \leq \left(\sum_{i\in[\alpha]} \frac{1}{i}\frac{q^{\frac{i}{2}}-1}{q-1}\right) + n$$
$$= \left(\sum_{i\in[\alpha]} \frac{1}{i}\frac{q(q^{\frac{i}{2}}-1)}{q(q-1)}\right) + n.$$

On the other hand, we have to compare this quantity with (4) which can be rewritten as

$$\sum_{i\in[\alpha]} \frac{1}{i}\frac{(q+1)^{\frac{i}{2}}-1}{q} = \sum_{i\in[\alpha]} \frac{1}{i}\frac{(q-1)((q+1)^{\frac{i}{2}}-1)}{q(q-1)}.$$

Thus the claim is proven, if the substraction of the latter one and the previous one is greater than zero, i.e., we show that

$$\left(\sum_{i\in[\alpha]} \frac{1}{i}\frac{(q-1)((q+1)^{\frac{i}{2}}-1)-q(q^{\frac{i}{2}}-1)}{q(q-1)}\right) - n > 0, \text{ i.e.} \tag{5}$$

$$\left(\sum_{i\in[\alpha]} \frac{1}{i}\frac{(q-1)(q+1)^{\frac{i}{2}}-qq^{\frac{i}{2}}+1}{q(q-1)}\right) - n > 0. \tag{6}$$

With $f(i) = \frac{(q-1)(q+1)^{\frac{i}{2}}-qq^{\frac{i}{2}}+1}{iq(q-1)}$ for all $i \in [\alpha]$, the proof of (6) contains the following steps

1. For all $i \geq 2$ we have $f(i) \geq 0$,
2. $f(5) + f(1) \geq 0$,
3. $f(\alpha) - n > 0$.

Step 1.: For $i = 2$ we have

$$f(2) = \frac{1}{2}\frac{(q-1)(q+1)-q^2+1}{q(q-1)} = \frac{q^2-1-q^2+1}{2q(q-1)} = 0.$$

For $i = 3$ we have

$$f(3) = \frac{1}{3}\frac{(q-1)(q+1)\sqrt{q+1}-q^2\sqrt{q}+1}{q(q-1)}.$$

Consider the function $g : \mathbb{R} \to \mathbb{R}; q \mapsto q^4 - 2q^3 - 2q^2 + q + 1$. This function has two minima (between -0.75 and -0.5 as well as between 1.75 and 2) and one maximum (between 0.125 and 0.25). Since g has only two inflexion points and g is strictly greater than zero at the first minima, g has only two roots. The first root is between 0.7 and 0.8 and the second root is between 2.5 and 2.75. Thus

for all $q \geq 2.75$ we have $g(q) > 0$. This implies $q^5 + q^4 - 2q^3 - 2q^2 + q + 1 > q^5$. Hence equivalently we get $(q+1)(q^4 - 2q^2 + 1) > q^5$, i.e., $(q+1)(q^2-1)^2 > qq^4$. This implies $\sqrt{q+1}(q^2-1) > \sqrt{q}q^2$ which proves that the numerator of $f(3)$ is positive and hence $f(3) > 0$. Before we prove the claim for $i \geq 4$, we will prove that $(q-1)(q+1)^j \geq q^{j+1}$ for $j \geq 2$. Firstly we get

$$(q-1)(q+1)^j = \left(\sum_{k \in [j]} \left(\binom{j}{k-1} - \binom{j}{k} \right) q^k \right) + q^{j+1} - 1.$$

Due to the central symmetry of each row of the Pascal triangle and since the distribution of the binomial coefficient is unimodal, for $k \leq \lfloor j/2 \rfloor$, we have

$$\binom{j}{j-k} - \binom{j}{j-k+1} = -\left(\binom{j}{k-1} - \binom{j}{k} \right) > 0$$

and thus

$$(q-1)(q+1)^j = \left(\sum_{k \in [\lfloor j/2 \rfloor]} \left(\binom{j}{k} - \binom{j}{k-1} \right) (q^{j-k+1} - q^k) \right) + q^{j+1} - 1.$$

Since $k \leq \lfloor j/2 \rfloor$, we have $j - k + 1 > k$ and each term of the above sum is thus positive. This shows that $(q-1)(q+1)^j \geq q^{j+1}$. This leads to the following estimations for $f(i)$. For $i = 2j$ and $j \geq 2$ we get

$$f(i) = \frac{(q-1)(q+1)^j - qq^j + 1}{iq(q-1)} \geq \frac{q^{j+1} - q^{j+1} + 1}{iq(q-1)} > 0.$$

Finally for $i = 2j+1$ and $j \geq 2$ we get

$$f(i) = \frac{(q-1)(q+1)^j \sqrt{q+1} - qq^j \sqrt{q} + 1}{iq(q-1)} \quad \geq \frac{q^{j+1}(\sqrt{q+1} - \sqrt{q}) + 1}{iq(q-1)} > 0.$$

Step 2.: Notice that $\alpha \geq 7$ holds and thus $f(5)$ is always a summand. For $f(5) + f(1)$ we have to prove

$$\frac{(q-1)(q+1)^2\sqrt{q+1} - qq^2\sqrt{q} + 1}{5q(q-1)} + \frac{(q-1)\sqrt{q+1} - q\sqrt{q} + 1}{q(q-1)} \geq 0$$

Thus we get for the numerator

$$\begin{aligned}
&(q-1)(q+1)^2\sqrt{q+1} - qq^2\sqrt{q} + 1 + 5(q-1)\sqrt{q+1} - 5q\sqrt{q} + 5 \\
&= (q-1)\sqrt{q+1}((q+1)^2 + 5) - q\sqrt{q}(q^2+5) + 6 \\
&= (q-1)\sqrt{q+1}(q^2 + 2q + 6) - q\sqrt{q}(q^2+5) + 6 \\
&= q^3\sqrt{q+1} + q^2\sqrt{q+1} + 4q\sqrt{q+1} - 6\sqrt{q+1} - q^3\sqrt{q} - 5q\sqrt{q} + 6.
\end{aligned}$$

We have $q^3\sqrt{q+1} > q^3\sqrt{q}$ and, since $q \geq 3$,

$$q^2\sqrt{q+1} \geq (6+q)\sqrt{q+1}.$$

Therefore $q^2\sqrt{q+1}+4q\sqrt{q+1} \geq 6\sqrt{q+1}+5q\sqrt{q}$ and the numerator is positive.

Step 3.: Notice that for fixed i, $f(i)$ is monotonically increasing for increasing q. This implies

$$f(\alpha) \geq \frac{2 \cdot 4^{\frac{\alpha}{2}} - 3 \cdot 3^{\frac{\alpha}{2}} + 1}{6\alpha} = \frac{2^{\alpha+1} - 3^{\frac{\alpha}{2}+1} + 1}{6\alpha}.$$

We are going to prove that

$$2^{\alpha+1} - 3^{\frac{\alpha}{2}+1} + 1 > 6\alpha n. \tag{7}$$

Recall that α is a function of n, given by $\alpha = \lfloor \frac{16}{7}\sqrt{n} \rfloor + 5$.

First, we have

$$2^{\alpha+1} - 3^{\frac{\alpha}{2}+1} > 2^{\alpha-1} - 2^{\frac{\alpha}{2}}.$$

Indeed, this inequality is equivalent to

$$2^{\frac{\alpha}{2}}\left(3 \cdot 2^{\frac{\alpha}{2}-1} + 1\right) > 3^{\frac{\alpha}{2}+1} \quad \Leftrightarrow \quad 2^{\frac{\alpha}{2}-1} + \frac{1}{3} > \left(\frac{3}{2}\right)^{\frac{\alpha}{2}}.$$

We can check that this last inequality is true by taking the logarithm of both sides, since $\alpha > 5$.

Therefore, it is sufficient for (7) to show that

$$2^{\alpha-1} - 2^{\frac{\alpha}{2}} = 2^{\frac{\alpha}{2}}(2^{\frac{\alpha}{2}-1} - 1) > 6\alpha n.$$

Note that $2^{\frac{\alpha}{2}} > n$ (indeed, $\lfloor \frac{16}{7}\sqrt{n} \rfloor + 5 > 2\sqrt{n}+5$, thus $2^{\frac{\alpha}{2}} > 2^{\frac{5}{2}} \cdot 2^{\sqrt{n}}$). Once again, taking the logarithms, one can check that $2^{\frac{5}{2}} \cdot 2^{\sqrt{n}} > n$ holds.

To verify (7), it remains to show that $2^{\frac{\alpha}{2}-1} - 1 \geq 6\alpha$ or that $2^{\frac{\alpha}{2}-1} > 6\alpha$. Taking the logarithms, it is equivalent to

$$\begin{aligned} &\frac{\alpha}{2} - 1 > \log(6) + \log(\alpha) \\ \Leftrightarrow &\alpha - 2\log(\alpha) > 2\log(6) + 2, \end{aligned}$$

which is true for $\alpha \geq 15$, that is for $n \geq 16$. Equation (7) can be verified by a computer for $q - 1 \leq n < 16$.

By 1., 2., and 3. Equation (6) is proven and this proves the claim. □