

Graph and String Parameters: Connections Between Pathwidth, Cutwidth and the Locality Number

Katrin Casel¹, Joel D. Day², Pamela Fleischmann³, Tomasz Kociumaka⁴,
Florin Manea³, and Markus L. Schmid⁵

¹Hasso Plattner Institute, University of Potsdam, Germany, Katrin.Casel@hpi.de

²Loughborough University, UK, J.Day@lboro.ac.uk

³Kiel University, Germany, {fpa,flm}@informatik.uni-kiel.de

⁴University of Warsaw, Poland, and Bar-Ilan University, Israel, kociumaka@mimuw.edu.pl

⁵Trier University, Germany, MLSchmid@MLSchmid.de

Abstract

We investigate the locality number, a recently introduced structural parameter for strings (with applications in pattern matching with variables), and its connection to two important graph-parameters, cutwidth and pathwidth. These connections allow us to show that computing the locality number is NP-hard but fixed parameter tractable (when the locality number or the alphabet size is treated as a parameter), and can be approximated with ratio $O(\sqrt{\log \text{opt}} \log n)$. As a by-product, we also relate cutwidth via the locality number to pathwidth, which is of independent interest, since it improves the currently best known approximation algorithm for cutwidth. In addition to these main results, we also consider the possibility of greedy-based approximation algorithms for the locality number.

1 Introduction

Graphs, on the one hand, and strings, on the other, are two different types of data objects and they have certain particularities. Graphs seem to be more popular in fields like classical and parameterised algorithms and complexity (due to the fact that many natural graph problems are intractable), while fields like formal languages, pattern matching, verification or compression are more concerned with strings. Moreover, both the field of graph algorithms as well as string algorithms are well established and provide rich toolboxes of algorithmic techniques, but they differ in that the former is tailored to computationally hard problems (e.g., the approach of treewidth and related parameters), while the latter focuses on providing efficient data-structures for near-linear-time algorithms. Nevertheless, it is sometimes possible to bridge this divide, i.e., by “flattening” a graph into a sequential form, or by “inflating” a string into a graph, to make use of respective algorithmic techniques otherwise not applicable. This paradigm shift may provide the necessary leverage for new algorithmic approaches.

In this paper, we are concerned with certain structural parameters (and the problems of computing them) for graphs and strings: the *cutwidth* $cw(G)$ of a graph G (i.e., the maximum number of “stacked” edges if the vertices of a graph are drawn on a straight line), the *pathwidth* $pw(G)$ of a graph G (i.e., the minimum width of a tree decomposition the tree structure of which is a path), and the *locality number* $\text{loc}(\alpha)$ of a string α (explained in more detail in the next paragraph). By CUTWIDTH, PATHWIDTH and LOC, we denote the corresponding decision problems and with the prefix MIN, we refer to the minimisation variants. The two former graph-parameters are very classical. Pathwidth is a simple (yet still hard to compute) subvariant of treewidth, which measures how much a graph resembles a path. The problems PATHWIDTH and MINPATHWIDTH are intensively studied (in terms of exact, parameterised and approximation algorithms) and have numerous applications (see the surveys and textbook [10, 35, 8]). CUTWIDTH is the best known example of a whole class of so-called *graph layout problems* (see the survey [17, 40] for detailed information), which are studied since the 1970s and were originally motivated by questions of circuit layouts.

The locality number is rather new and we shall discuss it in more detail. A word is k -local if there exists an order of its symbols such that, if we *mark* the symbols in the respective order (which is called a *marking sequence*), at each stage there are at most k contiguous blocks of marked symbols in the word. This k is called the *marking number* of that marking sequence. The *locality number* of a word is the smallest k for which that word is k -local, or, in other words, the minimum marking number over all marking sequences. For example, the marking sequence $\sigma = (x, y, z)$ marks $\alpha = xyxyzxz$ as follows (marked blocks are illustrated by overlines): $\overline{xy}xyzxz$, $x\overline{y}xyzxz$, $xy\overline{z}xyzxz$; thus, the marking number of σ is 3. In fact, all marking sequences for α have a marking number of 3, except (y, x, z) , for which it is 2: $x\overline{y}x\overline{y}zxxz$, $\overline{xy}x\overline{y}zxxz$, $\overline{xy}x\overline{y}zxxz$. Thus, the locality number of α , denoted by $\text{loc}(\alpha)$, is 2.

The locality number has applications in pattern matching with variables [14]. A *pattern* is a word that consists of *terminal symbols* (e.g., a, b, c), treated as constants, and *variables* (e.g., x_1, x_2, x_3, \dots). A pattern is mapped to a word by substituting the variables by strings of terminals. For example, $x_1x_1babx_2x_2$ can be mapped to $acacbabcc$ by the substitution ($x_1 \rightarrow ac, x_2 \rightarrow c$). Deciding whether a given pattern matches (i.e., can be mapped to) a given word is one of the most important problems that arise in the study of patterns with variables (note that the concept of patterns with variables arises in several different domains like combinatorics on words (word equations [31], unavoidable patterns [37]), pattern matching [1], language theory [2], learning theory [2, 19, 39, 43, 32, 22], database theory [7], as well as in practice, e.g., extended regular expressions with backreferences [26, 27, 45, 28], used in programming languages like Perl, Java, Python, etc.). Unfortunately, the *matching problem* is NP-complete [2] in general (it is also NP-complete for strongly restricted variants [23, 21] and also intractable in the parameterised setting [24]).

As demonstrated in [44], for the matching problem a paradigm shift as sketched in the first paragraph above yields a very promising algorithmic approach. More precisely, any class of patterns with bounded treewidth (for suitable graph representations) can be matched in polynomial-time. However, computing (and therefore algorithmically exploiting) the treewidth of a pattern is difficult (see the discussion in [21, 44]), which motivates more direct string-parameters that bound the treewidth and are simple to compute (virtually all known structural parameters that lead to tractability [14, 21, 44, 46] are of this kind (the efficiently matchable classes investigated in [15] are one of the rare exceptions)). This also establishes an interesting connection between ad-hoc string parameters and the more general (and much better studied) graph parameter treewidth. The locality number is a simple parameter directly defined on strings, it bounds the treewidth and the corresponding marking sequences can be seen as instructions for a dynamic programming algorithm. However, compared to other “tractability-parameters”, it seems to cover best the treewidth of a string, but whether it can be efficiently computed is unclear.

In this paper, we investigate the problem of computing the locality number and, by doing so, we establish an interesting connection to the graph parameters cutwidth and pathwidth with algorithmic implications for approximating cutwidth. In the following, we first discuss related results in more detail and then outline our respective contributions.

Known Results and Open Questions: For LOC, only exact exponential-time algorithms are known and whether it can be solved in polynomial-time, or whether it is at least fixed-parameter tractable is mentioned as open problems in [14]. Approximation algorithms have not yet been considered. Addressing these questions is the main purpose of this paper.

PATHWIDTH and CUTWIDTH are NP-complete, but fixed-parameter tractable with respect to parameter $\text{pw}(G)$ or $\text{cw}(G)$, respectively (even with “linear” fpt-time $g(k)O(n)$ [9, 11, 48]). With respect to approximation, their minimisation variants have received a lot of attention, mainly because they yield (like many other graph parameters) general algorithmic approaches for numerous graph problems, i.e., a good linear arrangement or path-decomposition can often be used for a dynamic programming (or even divide and conquer) algorithm. More generally speaking, pathwidth and cutwidth are related to the more fundamental concepts of small balanced vertex or edge separators for graphs (i.e., a small set of vertices (or edges, respectively) that, if removed, divides the graph into two parts of roughly the same size. More precisely, $\text{pw}(G)$ and $\text{cw}(G)$ are upper bounds for the smallest balanced *vertex* separator of G and the smallest balanced *edge* separator of G , respectively (see [20] for further details and explanations of the

algorithmic relevance of balanced separators). The best known approximation algorithms for MINPATHWIDTH and MINCUTWIDTH (with approximations ratios of $O(\sqrt{\log(\text{opt})} \log(n))$ and $O(\log^2(n))$, respectively) follow from approximations of vertex separators (see [20]) and edge separators (see [36]), respectively.

Our Contributions: There are two natural approaches to represent a word α over alphabet Σ as a graph $G_\alpha = (V_\alpha, E_\alpha)$: (1) $V_\alpha = \{1, 2, \dots, |\alpha|\}$ and the edges are somehow used to represent the actual symbols, or (2) $V_\alpha = \Sigma$ and the edges are somehow used to represent the positions of α . We present a reduction of type (2) such that $|E_\alpha| = O(|\alpha|)$ and $\text{cw}(G_\alpha) = 2 \text{loc}(\alpha)$, and a reduction of type (1) such that $|E_\alpha| = O(|\alpha|^2)$ and $\text{loc}(\alpha) \leq \text{pw}(G_\alpha) \leq 2 \text{loc}(\alpha)$. Since these reductions are parameterised reductions and also allow to transfer approximation results, we conclude that LOC is fixed-parameter tractable if parameterised by $|\Sigma|$ or by the locality number (answering the respective open problem from [14]), and also that there is a polynomial-time $O(\sqrt{\log(\text{opt})} \log(n))$ -approximation algorithm for MINLOC.

In addition, we also show a way to represent an arbitrary multi-graph $G = (V, E)$ by a word α_G over alphabet V , of length $|E|$ and with $\text{cw}(G) = \text{loc}(\alpha)$. This describes a Turing-reduction from CUTWIDTH to LOC which also allows to transfer approximation results between the minimisation variants. As a result, we can conclude that LOC is NP-complete (which solves the other open problem from [14]). Finally, by plugging together the reductions from MINCUTWIDTH to MINLOC and from MINLOC to MINPATHWIDTH, we obtain a reduction which transfers approximation results from MINPATHWIDTH to MINCUTWIDTH, which yields an $O(\sqrt{\log(\text{opt})} \log(n))$ -approximation algorithm for MINCUTWIDTH. This improves, to our knowledge for the first time since 1999, the best approximation for CUTWIDTH from [36].

To our knowledge, this connection between cutwidth and pathwidth has not yet been reported in the literature so far. This is rather surprising, since CUTWIDTH and PATHWIDTH have been jointly investigated in the context of exact and approximation algorithms, especially in terms of balanced vertex and edge separators. More precisely, the approximation of pathwidth and cutwidth follows from the approximation of vertex and edge separators, respectively, and the approximation of vertex separators usually relies on edge separators: the edge separator approximation from [36] can be used as a black-box for vertex separator approximation, and the best vertex separator algorithm from [20] uses a technique for computing edge separators from [4] as component. Our improvement, on the other hand, is achieved by going in the opposite direction: we use pathwidth approximation (following from [20]) in order to improve the currently best cutwidth approximation (from [36]). This might be why the reduction from cutwidth to pathwidth has been overlooked in the literature. Another reason might be that this relation is less obvious on the graph level and becomes more apparent if linked via the string parameter of locality, as in our considerations. Nevertheless, since pathwidth and cutwidth are such crucial parameters for graph algorithms, we also translate our locality based reduction into one from graphs to graphs directly.

Appendices A, B and C contain additional information and explanations on some results mentioned in this paper.

2 Preliminaries

Basic Definitions: The set of strings (or words) over an alphabet X is denoted by X^* , by $|\alpha|$ we denote the length of a word α , $\text{alph}(\alpha)$ is the smallest alphabet X with $\alpha \in X^*$. A string β is called a *factor* of α if $\alpha = \alpha' \beta \alpha''$; if $\alpha' = \varepsilon$ or $\alpha'' = \varepsilon$, where ε is the empty string, β is a *prefix* or a *suffix*, respectively. For a position j , $1 \leq j \leq |\alpha|$, we refer to the symbol at position j of α by the expression $\alpha[j]$, and $\alpha[j..j'] = \alpha[j] \alpha[j+1] \dots \alpha[j']$, $1 \leq j \leq j' \leq |\alpha|$. For a word α and $x \in \text{alph}(\alpha)$, let $\text{ps}_x(\alpha) = \{i \mid 1 \leq i \leq |\alpha|, \alpha[i] = x\}$ be the set of all positions where x occurs in α . For a word α , let $\alpha^0 = \varepsilon$ and $\alpha^{i+1} = \alpha \alpha^i$ for $i \geq 0$.

Let α be a word and let $X = \text{alph}(\alpha) = \{x_1, x_2, \dots, x_n\}$. A *marking sequence* is an enumeration, or ordering on the letters, and hence may be represented either as an ordered list of the letters or, equivalently, as a bijection $\sigma : \{1, 2, \dots, |X|\} \rightarrow X$. Given a word α and a marking sequence σ , the *marking number* $\pi_\sigma(\alpha)$ (of σ with respect to α) is the maximum number of marked blocks obtained while marking α according to σ . We say that α is k -local if and only if, for some marking sequence σ , we have $\pi_\sigma(\alpha) \leq k$, and the smallest k such that α is k -local

is the *locality number* of α , denoted by $\text{loc}(\alpha)$. A marking sequence σ with $\pi_\sigma(\alpha) = \text{loc}(\alpha)$ is *optimal* (for α). For a marking sequence $\sigma = (x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(m)})$ and a word α , by *stage i of σ* we denote the word α in which exactly the positions $\bigcup_{j=1}^i \text{ps}_{x_{\sigma(j)}}(\alpha)$ are marked. For a word α , the *condensed form of α* , denoted by $\text{cond}(\alpha)$, is obtained by replacing every maximal factor x^k with $x \in \text{alph}(\alpha)$ by x . For example, $\text{cond}(x_1x_1x_2x_2x_2x_1x_2x_2) = x_1x_2x_1x_2$. A word α is *condensed* if $\alpha = \text{cond}(\alpha)$.

Remark 1. For a word α , we have $\text{loc}(\text{cond}(\alpha)) = \text{loc}(\alpha)$ [14]. Hence, by computing $\text{cond}(\alpha)$ in time $O(|\alpha|)$, algorithms for computing the locality number (and the respective marking sequences) for *condensed* words extend to algorithms for general words.

Examples and Word Combinatorial Considerations: The structure of 1-local and 2-local words is characterised in [14]. The simplest 1-local words are repetitions x^k for some $k \geq 0$. Furthermore, if α is 1-local, then $y^\ell \alpha y^r$ is 1-local, where $y \notin \text{alph}(\alpha)$, $\ell, r \geq 0$. Marking sequences for 1-local words can be obtained by going from the “inner-most” letters to the “outer-most” ones. The English words *radar*, *refer*, *blender*, or *rotator* are all 1-local.

Generally, in order to have a high locality number, a word needs to contain many alternating occurrences of (at least) two letters. For instance, $(x_1x_2)^n$ is n -local. In general (see Appendix A), one can show that if $\text{loc}(w) = k$, then $\text{loc}(w^i) \in \{ik, ik - i + 1\}$.

The well-known *Zimin words* [37] also have high locality numbers compared to their lengths. These words are important in the domain of avoidability, as it was shown that a terminal-free pattern is unavoidable (i.e., it occurs in every infinite word over a large enough finite alphabet) if and only if it occurs in a Zimin word. The Zimin words Z_i , for $i \in \mathbb{N}$, are inductively defined by $Z_1 = x_1$ and $Z_{i+1} = Z_i x_{i+1} Z_i$. Clearly, $|Z_i| = 2^i - 1$ for all $i \in \mathbb{N}$. Regarding the locality of Z_i , note that marking x_2 leads to 2^{i-2} marked blocks; further, marking x_1 first and then the remaining symbols in an arbitrary order only extends or joins marked blocks. Thus, we obtain a sequence with marking number 2^{i-2} . In fact (see Appendix A), we have $\text{loc}(Z_i) = \frac{|Z_i|+1}{4} = 2^{i-2}$ for $i \in \mathbb{N}_{\geq 2}$. Notice that both Zimin words and 1-local words have an obvious palindromic structure. However, in the Zimin words the letters occur multiple times, but not in large blocks, while in 1-local words there are at most 2 blocks of each letter. One can show (see Appendix A) that if w is a palindrome, with $w = uau^R$ or $w = uu^R$, and $\text{loc}(u) = k$, then $\text{loc}(w) \in \{2k - 1, 2k, 2k + 1\}$ (u^R denotes the reversal of u).

The number of occurrences of a letter alone is not always a good indicator of the locality of a word. The German word *Einzelement* (basic component of a construction) has 5 occurrences of e , but is only 3-local, as witnessed by marking sequence (l, m, e, i, n, z, t) . Nevertheless, a repetitive structure often leads to high locality. The Finnish word *tutustuttu* (perfect passive of *tutustua*—to meet) is nearly a repetition and 4-local, while *pneumonoultramicroscopicsilicovolcanoconiosis* is an (English) 8-local word, and *lentokonesuihkuturbiinimoottoriapumekaanikkoaliupseerioppilas* is a 10-local (Finnish) word.

Complexity and Approximation: We briefly summarise the fundamentals of parameterised complexity [25, 18] and approximation [5]. A *parameterised problem* is a decision problem with instances (x, k) , where x is the actual input and $k \in \mathbb{N}$ is the *parameter*. A parameterised problem P is *fixed-parameter tractable* if there is an *fpt-algorithm* for it, i.e., one that solves P on input (x, k) in time $f(k) \cdot p(|x|)$ for a recursive function f and a polynomial p . We use the $O^*(\cdot)$ notation which hides multiplicative factors polynomial in $|x|$.

A minimisation problem P is a triple (I, S, m) with I being the *set of instances*, S being a function that maps instances $x \in I$ to the *set of feasible solutions* for x , and m being the *objective function* that maps pairs (x, y) with $x \in I$ and $y \in S(x)$ to a positive rational number. For every $x \in I$, we denote $m^*(x) = \min\{m(x, y) : y \in S(x)\}$. For $x \in I$ and $y \in S(x)$, the value $R(x, y) = \frac{m(x, y)}{m^*(x)}$ is the *performance ratio* of y with respect to x . An algorithm \mathcal{A} is an *approximation algorithm* for P with ratio $r : \mathbb{N} \rightarrow \mathbb{Q}$ (or an r -approximation algorithm, for short) if, for every $x \in I$, $\mathcal{A}(x) = y \in S(x)$, and $R(x, y) \leq r(|x|)$. We also let r be of the form $\mathbb{Q} \times \mathbb{N} \rightarrow \mathbb{Q}$ when the ratio r depends on $m^*(x)$ and $|x|$; in this case, we write $r(\text{opt}, |x|)$. We further assume that the function r is monotonically non-decreasing. Unless stated otherwise, all approximation algorithms run in polynomial time with respect to $|x|$.

Pathwidth, Cutwidth and Problem Definitions: Let $G = (V, E)$ be a (multi)graph with the vertices $V = \{v_1, \dots, v_n\}$. A *cut* of G is a partition (V_1, V_2) of V into two disjoint subsets

$V_1, V_2, V_1 \cup V_2 = V$; the (multi)set of edges $\mathcal{C}(V_1, V_2) = \{\{x, y\} \in E \mid x \in V_1, y \in V_2\}$ is called the cut-set or the (multi)set of edges crossing the cut, while V_1 and V_2 are called the sides of the cut. The *size* of this cut is the number of crossing edges, i.e., $|\mathcal{C}(V_1, V_2)|$. A *linear arrangement* of the (multi)graph G is a sequence $(v_{j_1}, v_{j_2}, \dots, v_{j_n})$, where (j_1, j_2, \dots, j_n) is a permutation of $(1, 2, \dots, n)$. For a linear arrangement $L = (v_{j_1}, v_{j_2}, \dots, v_{j_n})$, let $L(i) = \{v_{j_1}, v_{j_2}, \dots, v_{j_i}\}$. For every $i, 1 \leq i < n$, we consider the cut $(L(i), V \setminus L(i))$ of G , and denote the cut-set $\mathcal{C}_L(i) = \mathcal{C}(L(i), V \setminus L(i))$ (for technical reasons, we also set $\mathcal{C}_L(0) = \mathcal{C}_L(n) = \emptyset$). We define the *cutwidth* of L by $\text{cw}(L) = \max\{|\mathcal{C}_L(i)| \mid 0 \leq i \leq n\}$. Finally, the cutwidth of G is the minimum over all cutwidths of linear arrangements of G , i.e., $\text{cw}(G) = \min\{\text{cw}(L) \mid L \text{ is a linear arrangement for } G\}$.

A path decomposition (see [11]) of a connected graph $G = (V, E)$ is a tree decomposition whose underlying tree is a path, i.e., a sequence $Q = (B_0, B_1, \dots, B_m)$ (of *bags*) with $B_i \subseteq V, 0 \leq i \leq m$, satisfying the following two properties:

- *Cover property*: for every $\{u, v\} \in E$, there is an index $i, 0 \leq i \leq m$, with $\{u, v\} \subseteq B_i$.
- *Connectivity property*: for every $v \in V$, there exist indices i_v and $j_v, 0 \leq i_v \leq j_v \leq m$, such that $\{j \mid v \in B_j\} = \{i \mid i_v \leq i \leq j_v\}$. In other words, the bags that contain v occur on consecutive positions in (B_0, \dots, B_m) .

The *width* of a path decomposition Q is $w(Q) = \max\{|B_i| \mid 0 \leq i \leq m\} - 1$, and the *pathwidth* of a graph G is $\text{pw}(G) = \min\{w(Q) \mid Q \text{ is a path decomposition of } G\}$. A path decomposition is *nice* if $B_0 = B_m = \emptyset$ and, for every $i, 1 \leq i \leq m$, either $B_i = B_{i-1} \cup \{v\}$ or $B_i = B_{i-1} \setminus \{v\}$, for some $v \in V$.

It is convenient to treat a path decomposition Q as a scheme marking the vertices of the graph based on the order in which the bags occur in the bag sequence. More precisely, all vertices are initially marked as **open**. Then we process the bags one by one, as they occur in Q . When we process the first bag that contains a vertex v , then v becomes **active**. When we process the last bag that contains v , it becomes **closed**. The connectivity property enforces that vertices that are **closed** cannot be marked as **active** again, while the cover property enforces that adjacent vertices must be both **active** at some point. The width is the maximum number of vertices which are marked **active** at the same time minus one. If the path decomposition is nice, then whenever a bag is processed as described above, we change the marking of exactly one vertex.

We next formally define the computational problems of computing the parameters defined above. By LOC, CUTWIDTH and PATHWIDTH, we denote the problems to check for a given word α or graph G and integer $k \in \mathbb{N}$, whether $\text{loc}(\alpha) \leq k$, $\text{cw}(G) \leq k$, and $\text{pw}(G) \leq k$, respectively. Note that since we can assume that $k \leq |\alpha|$ and $k \leq |G|$, whether k is given in binary or unary has no impact on the complexity. With the prefix MIN, we refer to the minimisation variants. More precisely, $\text{MINLOC} = (I, S, m)$, where I is the set of words, $S(\alpha)$ is the set of all marking sequences for α and $m(\alpha, \sigma) = \pi_\sigma(\alpha)$ (note that $m^*(\alpha) = \text{loc}(\alpha)$); $\text{MINCUTWIDTH} = (I, S, m)$, where I are all multigraphs, $S(G)$ is the set of linear arrangements of G , and $m(G, L) = \text{cw}(L)$ (note that $m^*(G) = \text{cw}(G)$); finally, $\text{MINPATHWIDTH} = (I, S, m)$, where I are all graphs, $S(G)$ is the set of path decompositions of G , and $m(G, Q) = w(Q)$ (note that $m^*(G) = \text{pw}(G)$).

3 Locality and Cutwidth

In this section, we introduce polynomial-time reductions from LOC to CUTWIDTH and vice versa. The established close relationship between these two problems lets us derive several complexity-theoretic and algorithmic results for LOC. We also discuss approximation-preserving properties of our reductions.

First, we show a reduction from LOC to CUTWIDTH. For a word α and an integer $k \in \mathbb{N}$, we build a multigraph $H_{\alpha, k} = (V, E)$ whose set of nodes $V = \text{alph}(\alpha) \cup \{\$, \#\}$ consists of symbols occurring in α and two additional characters $\$, \# \notin \text{alph}(\alpha)$. The multiset of edges E contains an edge between nodes $x, y \in \text{alph}(\alpha)$ for each occurrence of the factors xy and yx in α , as well as $2k$ edges between $\$$ and $\#$, one edge between $\$$ and the first letter of α , and one edge between $\$$ and the last letter of α . An example is given in Figure 1.

Lemma 2. *The graph $H_{\alpha, k}$ satisfies $\text{cw}(H_{\alpha, k}) = 2k$ if and only if $\text{loc}(\alpha) \leq k$.*

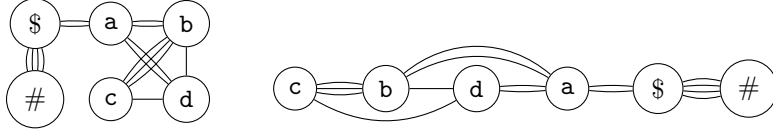


Figure 1: The graph $H_{\alpha,k}$ for $\alpha = abcdbcdbada$ and $k = 2$; an optimal linear arrangement of $H_{\alpha,k}$ with cutwidth 4 induces the optimal marking sequence (c, b, d, a) for α with marking number 2.

Proof. Suppose firstly that α is k -local, and let $\sigma = (x_1, x_2, \dots, x_n)$ be an optimal marking sequence of α . Consider the linear arrangement $L = (x_1, x_2, \dots, x_n, \$, \#)$. Clearly,

$$\begin{aligned} |\mathcal{C}(\{x_1, x_2, \dots, x_n, \$\}, \{\#\})| &= 2k \text{ and} \\ s|\mathcal{C}(\{x_1, x_2, \dots, x_n\}, \{\$, \#\})| &= 2. \end{aligned}$$

Now consider a cut $(K_1, K_2) = (\{x_1, x_2, \dots, x_i\}, \{x_{i+1}, \dots, x_n, \$, \#\})$ for $1 \leq i < n$. Every edge $e \in \mathcal{C}(K_1, K_2)$ is of the form $\{x_j, x_h\}$ with $j \leq i < h$, or of the form $\{\alpha[1], \$\}$ or $\{\$, \alpha[|\alpha|]\}$. Consequently, every edge $e \in \mathcal{C}(K_1, K_2)$ corresponds to a unique factor $x_j x_h$ or $x_h x_j$ of α with $j \leq i < h$ and, after exactly the symbols x_1, x_2, \dots, x_i are marked, x_j is marked and x_h is not, or to a unique factor $\alpha[1]$ or $\alpha[|\alpha|]$ and, after exactly the symbols x_1, x_2, \dots, x_i are marked, $\alpha[1]$ or $\alpha[|\alpha|]$ is marked. Since there can be at most k marked blocks in α after marking the symbols x_1, \dots, x_i , there are at most $2k$ such factors, which means that $|\mathcal{C}(K_1, K_2)| \leq 2k$. Thus $\text{cw}(H_{\alpha,k}) \leq 2k$. Note that any linear arrangement must at some point separate the nodes $\$$ and $\#$, meaning $\text{cw}(H_{\alpha,k}) \geq 2k$, so we get that $\text{cw}(H_{\alpha,k}) = 2k$.

Now suppose that the cutwidth of $H_{\alpha,k}$ is $2k$ and let L be an optimal linear arrangement witnessing this fact. Firstly, we note that L must either start with $\#$ followed by $\$$ (i.e., have the form $(\#, \$, \dots)$) or end with $\#$ preceded by $\$$ (i.e., have the form $(\dots, \$, \#)$). Otherwise, since $H_{\alpha,k}$ is connected, every cut separating $\$$ and $\#$ would be of size strictly greater than $2k$. Because a linear ordering and its mirror image have the same cutwidth, we may assume that the optimal linear arrangement has the form $L = (x_{\tau(1)}, x_{\tau(2)}, \dots, x_{\tau(n)}, \$, \#)$ for some permutation τ of $\{1, \dots, n\}$. Let σ be the marking sequence $(x_{\tau(1)}, x_{\tau(2)}, \dots, x_{\tau(n)})$ of α induced by τ . Suppose, for contradiction, that for some i , with $1 \leq i < n$, after marking $x_{\tau(1)}, \dots, x_{\tau(i)}$, we have $k' > k$ marked blocks. Furthermore, let $K_1 = \{x_{\tau(1)}, \dots, x_{\tau(i)}\}$ and $K_2 = \{x_{\tau(i+1)}, \dots, x_{\tau(n)}, \$, \#\}$. For every marked block $\alpha[s..t]$ that is not a prefix or a suffix of α , we have $\alpha[s], \alpha[t] \in K_1$ and $\alpha[s-1], \alpha[t+1] \in K_2$ and therefore $\{\alpha[s-1], \alpha[s]\}, \{\alpha[t], \alpha[t+1]\} \in \mathcal{C}(K_1, K_2)$. Moreover, for a marked prefix $\alpha[1..s]$, we have $\alpha[1], \alpha[s] \in K_1$ and $\$, \alpha[s+1] \in K_2$ and therefore $\{\alpha[1], \$\}, \{\alpha[s], \alpha[s+1]\} \in \mathcal{C}(K_1, K_2)$. Analogously, the existence of a marked suffix $\alpha[t..|\alpha|]$ leads to $\{\alpha[|\alpha|], \$\}, \{\alpha[t-1], \alpha[t]\} \in \mathcal{C}(K_1, K_2)$. Consequently, for each marked block we have two unique edges in $\mathcal{C}(K_1, K_2)$, which implies $|\mathcal{C}(K_1, K_2)| \geq 2k' > 2k$. This contradicts the assumption that L is a witness that $H_{\alpha,k}$ has cutwidth $2k$. Thus, α must be k -local as stipulated. \square

In the following, we briefly discuss the complexity of this reduction. Suppose we are given a word α and an integer $k \leq |\alpha|$. It is usual in string algorithmics to assume that α is over an integer alphabet, i.e., $\text{alph}(\alpha) \subseteq \{1, \dots, |\alpha|\}$. In this framework, the multigraph $H_{\alpha,k}$ can be constructed in $O(|\alpha|)$ time (e.g., represented as a list of vertices and a list of edges).

Lemma 3. *If there is an $r(\text{opt}, h)$ -approximation algorithm for MINCUTWIDTH running in $O(f(h))$ time for an input multigraph with h edges, then there is an $(r(2 \text{opt}, |\alpha|) + \frac{1}{\text{opt}})$ -approximation algorithm for MINLOC running in $O(f(|\alpha|) + |\alpha|)$ time on an input word α .*

Proof. As already indicated in the proof of Lemma 2, for $k = \text{loc}(\alpha)$, every linear arrangement for $H_{\alpha,k}$ naturally translates to a marking sequence for α . However, in an approximate linear arrangement, the vertices $\#$ and $\$$ do not have to be at the first (or last) position. Still, the marking sequence corresponding to the linear arrangement L can have not more than $\frac{\text{cw}(L)}{2} + 1$ blocks, since only suffix and prefix can be marked blocks which correspond to only one instead of two edges in a cut in $H_{\alpha,k}$. This observation remains valid, if we do not include the extra

vertices $\#$ and $\$$ in $H_{\alpha,k}$ in the reduction. Let H_α be the graph obtained from $H_{\alpha,k}$ (for some k) by removing the extra vertices $\#$ and $\$$ (observe that this also removes the dependence on k). Removing vertices only decreases the cutwidth, so Lemma 2 implies that $\text{cw}(H_\alpha) \leq 2m^*(\alpha)$. Let α be an instance of MINLOC and \mathcal{A} an $r(\text{opt}, h)$ -approximation for MINCUTWIDTH on multigraphs.

The approximation algorithm \mathcal{A} run on H_α returns a linear arrangement $L = \mathcal{A}(H_\alpha)$ with $\text{cw}(L) \leq r(\text{opt}, h) \text{cw}(H_\alpha)$. Let σ be the marking sequence corresponding to L , then $R(\alpha, \sigma) = \frac{\pi_\sigma(\alpha)}{m^*(\alpha)} \leq \frac{2}{\text{cw}(H_\alpha)} \left(\frac{\text{cw}(L)}{2} + 1 \right) = \frac{\text{cw}(L)}{\text{cw}(H_\alpha)} + \frac{1}{m^*(\alpha)} = R(H_\alpha, L) + \frac{1}{m^*(\alpha)}$. The performance ratio $R(H_\alpha, L)$ is at most $r(\text{opt}, h)$, where $h = |\alpha|$ is the number of edges in H_α . For the optimum value $k = m^*(\alpha)$, the cutwidth of $H_{\alpha,k}$ is at least $2k - 2$ and σ has performance ratio at most $r(2 \text{opt}, |\alpha|)$ (measured with respect to the optimum value k for MINLOC).

The overall approximation procedure builds the graph H_α in $O(|\Sigma|)$, runs \mathcal{A} on H_α in $O(f(|\alpha|))$ and translates the linear arrangement into a marking sequence σ in $O(|\Sigma|)$. This gives an $(r(2 \text{opt}, |\alpha|) + \frac{1}{\text{opt}})$ -approximation for MINLOC with running time in $O(f(|\alpha|) + |\alpha|)$. \square

For a reduction from CUTWIDTH to LOC, let $H = (V, E)$ be a connected multigraph, where V is the set of nodes and E the multiset of edges (for technical reasons, we assume $|V| \geq 2$). Let $H' = (V, E')$ be the multigraph obtained by duplicating every edge in H . As such, each node in H' has even degree, so there exists an Eulerian cycle C (i.e., a cycle visiting each edge exactly once) in H' , and, moreover, $\text{cw}(H') = 2 \text{cw}(H)$. For each edge $e \in E'$, let α_e be the word over V that corresponds to an arbitrary traversal of the Eulerian path P obtained from C by deleting e ; see Figure 2 for an example.

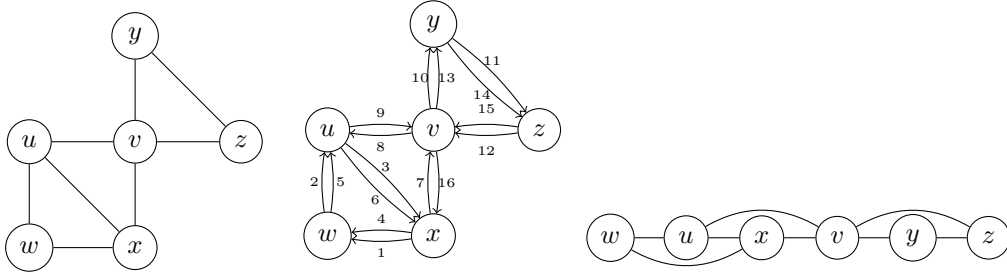


Figure 2: A graph H and its multigraph H' obtained by doubling the edges; the edge labels describe a Eulerian cycle that starts and ends in x . Deleting the edge (v, x) in this cycle yields the word $\alpha_{(v,x)} = xwu x w u x v v y z v y z v$, which has an optimal marking sequence (w, u, x, v, y, z) with marking number 3, and, thus, induces an optimal linear arrangement of H with cutwidth 3.

Lemma 4. *For any edge e in E' , the word α_e satisfies $\text{cw}(H) \leq \text{loc}(\alpha_e) \leq \text{cw}(H) + 1$. Moreover, there is a vertex $v \in V$ such that $\text{loc}(\alpha_e) = \text{cw}(H)$ for every edge e incident to v .*

Proof. Let $k = \text{cw}(H)$. Note that there is a natural bijection between the linear arrangements of H' and the marking sequences of the word α_e , since they both are essentially permutations of $\{1, 2, \dots, n\}$, i.e., for a permutation τ of $\{1, 2, \dots, n\}$, we can interpret $(x_{\tau(1)}, x_{\tau(2)}, \dots, x_{\tau(n)})$ both as the linear arrangement for H' and the marking sequence of α_e induced by τ . In the following, let τ be a permutation of $\{1, 2, \dots, n\}$, let $i \in \{1, 2, \dots, n-1\}$, $K_1 = \{x_{\tau(1)}, x_{\tau(2)}, \dots, x_{\tau(i)}\}$ and $K_2 = \{x_{\tau(i+1)}, \dots, x_{\tau(n)}\}$, and let $\mathcal{C}(K_1, K_2) = 2\ell$ (note that since every edge has been duplicated, we can guarantee that the size of every cut of H' is even).

Now consider α_e after after marking the letters $x_1, \dots, x_{\tau(i)}$. For every marked block $\alpha[s..t]$ that is not a prefix or a suffix of α , we have $\alpha[s], \alpha[t] \in K_1$ and $\alpha[s-1], \alpha[t+1] \in K_2$ and therefore $\{\alpha[s-1], \alpha[s]\}, \{\alpha[t], \alpha[t+1]\} \in \mathcal{C}(K_1, K_2)$. Moreover, for a marked prefix $\alpha[1..s]$, we have $\alpha[s] \in K_1$ and $\alpha[s+1] \in K_2$ and therefore $\{\alpha[s], \alpha[s+1]\} \in \mathcal{C}(K_1, K_2)$. Analogously, the existence of a marked suffix $\alpha[t..|\alpha|]$ leads to $\{\alpha[t-1], \alpha[t]\} \in \mathcal{C}(K_1, K_2)$.

Conversely, for every edge in $\mathcal{C}(K_1, K_2)$, with the exception of e (if e is in $\mathcal{C}(K_1, K_2)$ at all), there is a unique factor $\alpha_e[p..p+1]$ of α_e such that either $\alpha_e[p]$ is marked and $\alpha_e[p+1]$ is

unmarked, or vice-versa. Thus, if all marked blocks are internal, i.e., no marked block is a prefix or a suffix, then there are exactly ℓ marked blocks. Also, if both a prefix and a suffix occurs as a marked block, then we have $\ell + 1$ marked blocks. Finally, if a prefix occurs as a marked block, but no suffix, or vice-versa, then there are only ℓ marked blocks; note that in this case we must have $e \in \mathcal{C}(K_1, K_2)$. Since we consider all permutations, the arguments above are sufficient to conclude that, in our setting, each α_e has locality number either k or $k + 1$.

Furthermore, consider a linear ordering $L = (x_{j_1}, \dots, x_{j_n})$ of H' which is optimal, i.e., $|\mathcal{C}_L(i)| \leq 2k$. Note that if either the first or last letter of α_e is the last letter x_{j_n} to be marked according to the marking sequence induced by the linear ordering $(x_{j_1}, \dots, x_{j_n})$, the case that both a suffix and prefix of α_e are marked cannot be reached until $i = n$ and the entire word is marked. Consequently, this would imply that α_e has locality number k . For any permutation of the linear ordering $(x_{j_1}, \dots, x_{j_n})$, this holds for α_e where e is an edge adjacent to the node x_{j_n} , since the path P obtained by removing such an edge e from C must start or end with x_{j_n} . \square

The resulting Turing reduction from CUTWIDTH to LOC is performed in $O(nh)$ time, where $n = |V|$ is the number of vertices and $h = |E|$ is the number of edges of the input multigraph: First, the graph H' and its Eulerian cycle are constructed in $O(h)$ time. Then, for each vertex, we select an arbitrary incident edge e and build the word α_e of length $O(h)$.

Lemma 5. *If there is an $r(\text{opt}, |\alpha|)$ -approximation algorithm for MINLOC running in $O(f(|\alpha|))$ time on a word α , then there is an $r(\text{opt}, h)$ -approximation algorithm for MINCUTWIDTH running in $O(n(f(h) + h))$ time on a multigraph with n vertices and h edges.*

Proof. Let $G = (V, E)$ be an instance of MINCUTWIDTH and \mathcal{A} an $r(\text{opt}, |\alpha|)$ -approximation for MINLOC. By Lemma 4, there exists a vertex $v \in V$ such that $\text{loc}(\alpha_e) = \text{cw}(G)$ holds for any edge $e \in E$ adjacent to v . The approximation algorithm \mathcal{A} hence returns on input α_e a marking sequence σ with $\pi_\sigma(\alpha_e) \leq r(\text{opt}, |\alpha|) \text{cw}(G)$.

In the proof of Lemma 4 it is further shown that any marking sequence σ for α_e translates to a linear arrangement L for G with $\text{cw}(L) \leq \pi_\sigma(\alpha_e)$. The performance ratio of this linear arrangement is $R(G, L) = \frac{\text{cw}(L)}{\text{cw}(G)} \leq \frac{\pi_\sigma(\alpha_e)}{\text{loc}(\alpha_e)} \leq R(\alpha_e, \sigma)$.

The procedure which, for each vertex $v \in V$, constructs α_e for some $e \in E$ adjacent to v in $O(h)$, runs \mathcal{A} in $O(f(|\alpha_e|)) = O(f(h))$ and checks the resulting linear arrangement in $O(h)$ and returns the best linear arrangement among all $v \in V$, yields an $r(\text{opt}, h)$ -approximation for MINCUTWIDTH on multigraphs in $O(n(f(h) + h))$. \square

Consequences: In the following, we overview a series of complexity-theoretic and algorithmic consequences of the reductions provided above. We first discuss negative results and note that we can close one of the main problems left open in [14].

Theorem 6. *The LOC problem is NP-complete.*

Proof. Lemma 4 shows a polynomial time Turing reduction from CUTWIDTH to LOC. Indeed, given a (multi)graph H we construct in linear time the multigraph H' by duplicating its edges. H' has an Eulerian cycle, so, using Hierholzer's algorithm, we can compute such a cycle in linear time [30]. Let C be the computed Eulerian cycle. For each edge e of C construct, in linear time, the word α_e as described before Lemma 4. By Lemma 4 we get that $\text{cw}(H) = \frac{\text{cw}(H')}{2} = \min\{\text{loc}(\alpha_e) \mid e \text{ edge of } C\}$. This completes the reduction, and, thus, as CUTWIDTH is NP-hard (see, e.g., [17]), we get that loc is also NP-hard. As LOC clearly belongs to NP, the result follows. \square

Theorem 6 follows from the Turing reduction from CUTWIDTH to LOC, but it can also be proved using a polynomial-time one-to-many reduction from the well known NP-complete problem CLIQUE. This alternative approach (given in Appendix B) is more technically involved, but has the merit of emphasising how the combinatorial properties of the locality number can be used to construct computationally hard instances of LOC. Moreover, by the word-combinatorial observations about locality made in Section 2, it is clear that LOC is NP-complete also for words with special structure, e.g., palindromes and repetitions.

With respect to approximation, it is known that, assuming the Small Set Expansion Conjecture (denoted SSE; see [41]), there exists no constant-ratio approximation for MINCUTWIDTH

(see [49]). Consequently, approximating MINLOC within any constant factor is also SSE-hard. In particular, we point out that stronger inapproximability results for MINCUTWIDTH are not known. Positive approximation results for MINLOC will be discussed in Section 4.

On certain graph classes, the SSE conjecture is equivalent to the Unique Games Conjecture [33] (see [41, 42]), which, at its turn, was used to show that many approximation algorithms are tight [34] and is considered a major conjecture in inapproximability. However, some works seem to provide evidence that could lead to a refutation of SSE; see [3, 6, 29]. In this context, we show in Section 4 a series of unconditional results which state that multiple natural greedy strategies do not provide low-ratio approximation of MINLOC.

As formally stated next, Lemma 2 extends algorithmic results for computing cutwidth to determining the locality number (we formulate this result so that it also covers fpt-algorithms with respect to the standard parameters $\text{cw}(G)$ and $\text{loc}(\alpha)$). Note that the maximum degree in a multigraph G is bounded from above by $2\text{cw}(G)$, so the number of nodes n and the number of edges h satisfy $h \leq n \cdot \text{cw}(G)$. Hence, we state the complexity in terms of n and $\text{cw}(G)$ rather than with respect to h , which is the actual input size.

Lemma 7. *If there is an algorithm solving MINCUTWIDTH (resp., CUTWIDTH) in $O(f(\text{cw}(G), n))$ time for a multigraph G with n vertices, then there is an algorithm solving MINLOC (resp., LOC) in $O(f(2\text{loc}(\alpha), |\Sigma| + 2) + |\alpha|)$ time for a word α over an alphabet Σ .*

Proof. We only show the claim for MINCUTWIDTH; the case of CUTWIDTH follows immediately from 2. Our goal is to compute $\text{loc}(\alpha)$ for the word α , i.e., the minimum k such that α is k -local. By Lemma 2, we get $\text{cw}(H_{\alpha,k}) = 2k$ for $k \geq \text{loc}(\alpha)$ and $\text{cw}(H_{\alpha,k}) > 2k$ for $k < \text{loc}(\alpha)$. Consider a multigraph H_α obtained by removing the vertices $\#$ and $\$$ from $H_{\alpha,i}$ (the result does not depend on $i \in \mathbb{N}$), and observe that $2\text{loc}(\alpha) - 4 \leq \text{cw}(H_\alpha) \leq 2\text{loc}(\alpha)$. Indeed, if $\text{cw}(H_\alpha) < 2\text{loc}(\alpha) - 4$, we add the two missing nodes $\#$ and $\$$ (in this order) as a prefix to an optimal linear arrangement for H_α and get a linear arrangement of $H_{\alpha, \text{loc}(\alpha)-1}$ of width $2\text{loc}(\alpha) - 2$, a contradiction.

Hence, in order to determine $\text{loc}(\alpha)$, we proceed as follows: Compute $\ell = \text{cw}(H_\alpha)$ and iterate over integers k , $\frac{\ell}{2} \leq k \leq \frac{\ell+4}{2}$, in the increasing order, checking if $\text{cw}(H_{\alpha,k}) = 2k$. The first value for which this equality holds equals $\text{loc}(\alpha)$, and the marking sequence induced by the respective linear arrangement of $H_{\alpha,k}$ is an optimal one for α (as proved in 2). \square

In particular, we can draw the following corollaries using Lemma 7 and known results from the literature. Due to the algorithms of [12], which also work for multigraphs¹, MINLOC can be solved in $O^*(2^{|\Sigma|})$ time and space, or in $O^*(4^{|\Sigma|})$ time and polynomial space. In particular, this also implies that LOC is fixed-parameter tractable with respect to the alphabet size. Moreover, the fpt-algorithm from [48] directly implies that MINLOC is fixed-parameter tractable for parameter $\text{loc}(\alpha)$ with linear fpt-running-time $g(\text{loc}(\alpha))O(n)$. Since CUTWIDTH is NP-complete already for graphs with maximum degree 3 (see [38]), we also derive a stronger statement compared to Theorem 6: LOC is NP-complete even if every symbol has at most 3 occurrences; if every symbol has at most 2 occurrences, the complexity of LOC is open, while the case where every symbol has only one occurrence is trivial. If, on the other hand, the symbols have many occurrences in comparison to $|\alpha|$, i.e., $|\Sigma| = O(\log(|\alpha|))$, then LOC can be solved in polynomial time, e.g., using the $O^*(2^{|\Sigma|})$ -time algorithm mentioned above.

4 Locality and Pathwidth

In this section, we consider the approximability of the minimisation problem MINLOC. Since a marking sequence is just a linear arrangement of the symbols of the input word, this problem seems to be well tailored to greedy algorithms: until all symbols are marked, we choose an unmarked symbol according to some greedy strategy and mark it. There are two aspects that motivate the investigation of such approaches. Firstly, ruling out simple strategies is a natural initial step in the search for approximation algorithms for a new problem. Secondly, due to the results of Section 3, the obvious greedy approaches for computing the locality number may also

¹These algorithms actually support weighted graphs without any major modification and in the same complexity. In this setting, parallel edges connecting two vertices are replaced by a single “super-edge” whose weight is the number of parallel edges.

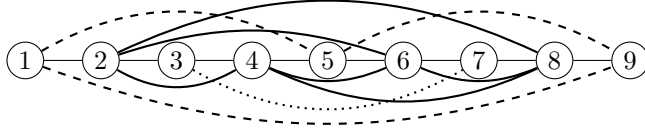


Figure 3: The graph G_α for $\alpha = \text{cabacabac}$; the three cliques are drawn with different edge-types.

provide a new angle to approximating the cutwidth of a graph, i.e., some greedy strategies may only become apparent in the locality number point of view and hard to see in the graph formulation of the problem. Given the fact that, as formally stated later as Theorem 14, approximating the cutwidth via approximation of the locality number does in fact improve the best currently known cutwidth approximation ratio, this seems to be a rather important aspect.

Unfortunately, we can formally show that many natural candidates for greedy strategies fail to yield promising approximation algorithms (and are therefore also not helpful for cutwidth approximation). We just briefly mention these negative results; all details are provided in Appendix C. The four considered basic strategies are the following: (1) prefer symbols with few occurrences, (2) symbols with many occurrences, (3) symbols leading to fewer blocks after marking, (4) symbols with earlier leftmost occurrence. All these strategies fail in a sense that there are arbitrarily long (condensed) words α with constant locality numbers for which these strategies yield marking sequences with marking numbers $\Omega(|\alpha|)$.

A more promising approach is to choose among symbols that extend at least one already marked block (except when marking the first symbol). We denote this strategy by **BlockExt** and marking sequences that can be obtained by it are called **BlockExt**-marking sequences. Intuitively, marking a symbol that has only isolated occurrences, and therefore will increase the current number of marked blocks by the number of its occurrences, seems a bad choice. This raises a general question whether every word has a **BlockExt**-marking sequence that is also optimal for this word. We answer this question negatively: all **BlockExt**-marking sequences for words like $x_1yx_2yx_3y \dots x_{2k}y$ achieve a marking number of at least $2k - 1$, while first marking x_2, x_3, \dots, x_{k+1} in this order (which all have only isolated occurrences), then y , and then the rest of the symbols in some order, yields at most k marked blocks. However, this only shows a lower bound of roughly 2 for the approximation ratio of algorithms based on **BlockExt**, so **BlockExt** might still be a promising candidate. However, in order to devise a **BlockExt**-based approximation algorithm, we still face the problem of deciding which of the extending symbols should be chosen; trying out all of them is obviously too costly. Unfortunately, if we handle this decision by one of the basic strategies (1)–(4) from above, e.g., choosing among all extending symbols one that leads to fewer new blocks, we again end up with poor approximation ratios. More precisely, we can again find arbitrarily long words α with constant locality numbers for which these algorithms yield marking numbers $\Omega(|\alpha|)$. Moreover, this is also true if we choose among all extending symbols one that has a maximum number of extending occurrences or one that maximises the ratio $\frac{\#\text{extending occ.}}{\#\text{occ.}}$.

While we obviously have not investigated *all* reasonable greedy strategies, we consider our negative results as sufficient evidence that a worthwhile approximation algorithm for computing the locality number most likely does not follow from such simple greedy strategies.

In the following, we adopt a more sophisticated approach of approximating the locality number: we devise a reduction to the problem of computing the pathwidth of a graph. To this end, we first have to describe how a (condensed) word can be represented as a graph: For a condensed word α , the graph $G_\alpha = (V_\alpha, E_\alpha)$ is defined by $V_\alpha = \{1, 2, \dots, |\alpha|\}$ and $E_\alpha = \{\{i, i + 1\} \mid 1 \leq i \leq |\alpha| - 1\} \cup \{\{i, j\} \mid \{i, j\} \subseteq \text{ps}_x(\alpha) \text{ for some } x \in \text{alph}(\alpha)\}$. Intuitively, G_α is obtained by interpreting every position of α as a vertex, connecting neighbouring positions by edges, and turning every set $\text{ps}_x(\alpha)$, $x \in \text{alph}(\alpha)$, into a clique (see Figure 3).

We use G_α as a unique graph representation for condensed words and whenever we talk about a path decomposition for α , we actually refer to a path decomposition of G_α and, since G_α has the positions of α as its vertices, the marking scheme behind a path decomposition (and its respective terminology) directly translates to a marking scheme of the positions of α .

Lemma 8. *Let α be a condensed word. Then $\text{pw}(G_\alpha) \leq 2 \text{loc}(\alpha)$.*

Proof. Let $\sigma = (x_1, x_2, \dots, x_m)$ be a marking sequence for α with $\pi_\sigma(\alpha) = k$. We describe a path-decomposition Q for G_α as a marking. First, for every i , $1 \leq i \leq m$, we define the step p_i of Q as the following situation. Every position that is a border position of a marked block at step i of σ is **active**, every other position that is marked at step i of σ is **closed**, and all other positions are **open**. Intuitively speaking, step p_i represents the marked factors of step i of σ in a natural way. The path-decomposition produces these steps in the order p_1, p_2, \dots, p_m . The step p_1 is obtained from the initial one (i. e., where all positions are **open**) by just setting all positions of $\text{ps}_{x_1}(\alpha)$ to **active**. The final step of Q where all positions are **closed** is obtained from step p_m by setting the only **active** positions 1 and $|\alpha|$ to **closed**. In the following, we describe how we reach p_{i+1} from p_i for every i with $1 \leq i \leq m - 1$.

Let s be arbitrary with $1 \leq s \leq m - 1$. In order to produce step p_{s+1} of Q from step p_s , we do the following:

1. For all $j \in \text{ps}_{x_{s+1}}(\alpha)$
 - (a) If marking j does not produce a new marked block of size 1
 - i. Set j to **active**
 - ii. If $j - 1$ is **active**, $j > 2$ and $j - 2$ is not open, then set $j - 1$ to **closed**
 - iii. If $j + 1$ is **active**, $j < |\alpha| - 1$ and $j + 2$ is not open, then set $j + 1$ to **closed**
2. Set all remaining positions from $\text{ps}_{x_{s+1}}$ to **active**
3. Set all positions from $p_{x_{s+1}}$ having only **active** or **closed** neighbours to **closed**.

Note that if in Step 1a we have $j = 2$ and $j - 1 = 1$ is **active**, it will remain **active**. Similarly, if $j = |\alpha| - 1$ and $j + 1 = |\alpha|$ is **active**, it will remain **active**. It can be easily seen, that after Step 3 is finished, we have reached step p_{s+1} of Q . Consequently, we have now fully defined Q . We note that when Step 1 is finished, then all positions of $\text{ps}_{x_{s+1}}$ that do not create new marked blocks are **active**, and we denote this step of Q as step p'_s . Moreover, we denote the situation reached after Step 2 is finished as step p''_s of Q .

In order to see that Q is a valid path-decomposition, we first observe that for every i , $1 \leq i \leq m$, we reach a step where all positions of $\text{ps}_{x_i}(\alpha)$ are **active** (namely step p_1 , if $i = 1$ and step p''_{i-1} otherwise), and for every j , $1 \leq j \leq |\alpha| - 1$, j is set from **active** to **closed** while $j + 1$ is **active**, or $j + 1$ is set from **active** to **closed** while j is **active**. Thus, Q satisfies the cover-property (the connectivity-property is trivially satisfied, since we define Q as a marking) and therefore is a valid path-decomposition. It remains to determine the width of Q .

Let s be arbitrary with $1 \leq s \leq m - 1$, let k_s and k_{s+1} be the number of marked blocks at steps s and $s + 1$ of σ , respectively. We note that $k_i \leq k$ and $k_{i+1} \leq k$. Now let us assume that in going from step s to step $s + 1$ of σ , exactly q new marked blocks of size 1 are created and r times we join a marked block with another marked block. This means that in step p'_s , we have the r **active** positions from $\text{ps}_{x_{s+1}}(\alpha)$ that are responsible for joining marked blocks of step s of σ , and in addition to that, for every marked block of step $s + 1$ of σ that is not a *new* block of size 1, we have at most 2 **active** border positions (these might or might not be from $\text{ps}_{x_{s+1}}(\alpha)$). Thus, there are at most $r + 2(k_{s+1} - q)$ **active** positions in step p'_s . Moreover, in step p'_s , we get an additional number of q **active** positions from $\text{ps}_{x_{s+1}}(\alpha)$ for the new marked blocks of size 1. In total, this leads to $r + 2(k_{s+1} - q) + q = 2k_{s+1} + r - q$ **active** positions in step p'_s . Finally, in reaching step p_{s+1} from p'_s , the number of **active** positions can only decrease.

It can be easily seen that, according to how Q is defined above, in going from p_s to p'_s , the number of **active** positions is always bounded by $\ell + 1$, where ℓ is the number of **active** positions in step p_s . Analogously, in going from p'_s to p''_s , the number of **active** positions is always bounded by $\ell' + 1$, where ℓ' is the number of **active** positions in step p'_s . We conclude that the number of **active** position in all the steps between p_s and p_{s+1} is bounded by $\max(2k_s, 2k_{s+1} + r - q) + 1$. Since we obviously have $k_{s+1} = k_s + q - r$, we get that

$$\max(2k_s, 2k_{s+1} + r - q) + 1 = \max(2k_s, k_s + k_{s+1}) + 1 \leq 2k + 1.$$

Obviously, the number of **active** positions in step p_1 and the steps preceding it is at most k . Therefore, $\text{pw}(Q) \leq 2k$, and therefore also $\text{pw}(G_\alpha) \leq 2 \text{loc}(\alpha)$. \square

Lemma 9. *Let α be a condensed word with $|\alpha| \geq 2$. Then $\text{loc}(\alpha) \leq \text{pw}(G_\alpha)$.*

Proof. Let $Q = (B_0, B_1, B_2, \dots, B_{2|\alpha|})$ be an arbitrary nice path-decomposition for G_α . For every i , $1 \leq i \leq m$, let p_i be the first step of Q where all positions of $\text{ps}_{x_i}(\alpha)$ are **active**. Without loss of generality, we assume that $p_1 < p_2 < \dots < p_m$. Let $\sigma = (x_1, x_2, \dots, x_m)$ and let $k = \pi_\sigma(\alpha)$. We now prove that one of the following cases hold:

- There is a step of Q with at least $k + 1$ **active** positions.
- There is a step of Q with at least k **active** positions and a marking sequence σ' with $\pi_{\sigma'}(\alpha) = k - 1$.

This implies that, for every path-decomposition Q of G_α , $\text{loc}(\alpha) \leq \text{pw}(Q)$ and therefore also $\text{loc}(\alpha) \leq \text{pw}(G_\alpha)$.

Let s , $1 \leq s \leq m$, be chosen such that the maximum number of marked blocks in α according to σ is reached for the first time at step s . In the following, we represent the marked version of α at step s of σ as a word $\hat{\alpha}$ over the set of symbols $\{\mathbf{o}, \mathbf{a}, \mathbf{c}\}$ which indicate the status of the positions at step p_s of Q . More formally, $\hat{\alpha}[i] = \mathbf{o}$ if position i is **open**, $\hat{\alpha}[i] = \mathbf{a}$ if position i is **active** and $\hat{\alpha}[i] = \mathbf{c}$ if position i is **closed** at step p_s of Q . Moreover, we consider the factorisation

$$\hat{\alpha} = \beta_0 \mu_1 \beta_1 \mu_2 \dots \mu_k \beta_k,$$

where the factors β_i , $0 \leq i \leq k$, correspond to the unmarked regions of α , and μ_i , $1 \leq i \leq k$, correspond to the marked blocks. Next, we establish some simple properties of $\hat{\alpha}$ that all follow directly from the definitions.

Obviously, $\beta_0, \beta_k \in \{\mathbf{a}, \mathbf{o}\}^*$, while $\beta_i \in \{\mathbf{a}, \mathbf{o}\}^+$ and $\mu_i \in \{\mathbf{a}, \mathbf{c}\}^+$ for every i with $1 \leq i \leq k$. This follows from the fact that an occurrence of a symbol y is **closed** at step p_s of Q if and only if it has already been marked before, i. e., $y \in \{x_1, x_2, \dots, x_{s-1}\}$. Moreover, for every i , $1 \leq i \leq k-1$, if $\mu_i[[\mu_i]] = \mathbf{c}$, then $\beta_i[1] = \mathbf{a}$, since otherwise, there is an **closed** position adjacent to an **open** one, which is a contradiction, since these two positions are also adjacent in G_α . For μ_k , this only holds if $\beta_k \neq \varepsilon$. An analogous observation can be made with respect to the leftmost positions of the factors μ_i , $1 \leq i \leq k$. Consequently, the first (or last) position of every marked block (that is not a prefix, or not a suffix, respectively) is at step p_s of Q either **active** or it is **closed** and preceded (or followed, respectively) by an **active** position.

We note further that all occurrences of x_s are contained in marked blocks at step s of σ and **active** at step p_s of Q , i. e., they all correspond to occurrences of \mathbf{a} in factors μ_i . Moreover, there is at least one occurrence of x_s , i. e., a position j with $\alpha[j] = x_s$. In the following, we assume that this position is in μ_r for some r with $1 \leq r \leq k$, i. e., $|\beta_0 \mu_1 \dots \beta_{r-1}| + 1 \leq j \leq |\beta_0 \mu_1 \dots \beta_{r-1} \mu_r|$ and $\mu_r = \nu_1 \mathbf{a} \nu_2$ with $|\beta_0 \mu_1 \dots \beta_{r-1} \nu_1| + 1 = j$.

Next, to every marked block μ_i , $1 \leq i \leq k$, we allocate a distinct position t_i that is **active** at step p_s of Q . First, we set $t_r = j$, i. e., the occurrence of x_s in the marked block μ_r . For every i , $1 \leq i < r$, we let t_i be an **active** position in μ_i , if one exists and $t_i = |\beta_0 \mu_1 \dots \beta_{i-1} \mu_i| + 1$ otherwise. Note that if μ_i does not contain any **active** position, then its rightmost occurrence is **closed** and therefore, as observed above, $\beta_i[1]$ is in fact **active**. We proceed analogously for the remaining marked blocks, i. e., for every i , $r < i \leq k$, we let t_i be some **active** position in μ_i , if one exists and $t_i = |\beta_0 \mu_1 \dots \beta_{i-1}|$ otherwise. Since every t_i with $1 \leq i < r$ is in $\mu_i \beta_i[1]$, every t_i with $r < i \leq k$ is in $\beta_{i-1}[[\beta_i]] \mu_i$, and t_r is in μ_r , these positions t_i are in fact k distinct positions that are **active** at step p_s of Q .

Now, if there is at least one additional **active** position, then there are at least $k + 1$ **active** positions at step p_s of Q , which implies that the first of the two cases from the beginning of the proof holds. So in the following, we assume that the **active** positions t_i , $1 \leq i \leq k$, are the only **active** positions at step p_s of Q .

In the following, we have to consider several cases. To this end, it makes sense to divide $\hat{\alpha}$ into the part left of μ_r , the factor μ_r and the part right of μ_r ; in particular, all our following observations for the left part shall also hold analogously for the right part. More precisely, we set $\hat{\alpha}_1 = \beta_0 \mu_1 \beta_1 \mu_2 \dots \beta_{r-1}$ (which we shall call the *left side*) and $\hat{\alpha}_2 = \beta_r \mu_{r+1} \beta_{r+1} \dots \mu_k \beta_k$ (which we shall call the *right side*), i. e., $\hat{\alpha} = \hat{\alpha}_1 \mu_r \hat{\alpha}_2$.

Now we take a closer look at the left side $\hat{\alpha}_1$. If, for some ℓ , $1 \leq \ell < r$, t_ℓ is not in μ_ℓ , then it corresponds to the leftmost position of β_ℓ . Moreover, we then have $\mu_\ell \in \{\mathbf{c}\}^+$, which implies

that the rightmost occurrence of $\beta_{\ell-1}$ must be **a**, which means that this is actually the position $t_{\ell-1}$ leading to $\beta_{\ell-1} = \mathbf{a}$. This argument can then be repeated, which means that if, for some ℓ , $1 \leq \ell < r$, t_ℓ is not in μ_ℓ , then $\beta_i = \mathbf{a}$ for every i with $1 \leq i \leq \ell - 1$, while for β_ℓ the leftmost position if **active**. In particular, this also means that $\mu_i \in \{\mathbf{c}\}^+$ for every i with $1 \leq i \leq \ell$, and that $\beta_0 = \varepsilon$. This can be illustrated as follows:

$$\widehat{\alpha}_1 = \mu_1 \underbrace{\mathbf{a}}_{\beta_1} \mu_2 \underbrace{\mathbf{a}}_{\beta_2} \cdots \mu_{\ell-1} \underbrace{\mathbf{a}}_{\beta_{\ell-1}} \mu_\ell \underbrace{\mathbf{a} \mathbf{o}^{g_1}}_{\beta_\ell} \mu_{\ell+1} \beta_{\ell+1} \mu_{\ell+2} \cdots \beta_{r-2} \mu_{r-1} \beta_{r-1},$$

where $\mu_i \in \{\mathbf{c}\}^+$ for every i with $1 \leq i \leq \ell$, and $g_1 \geq 0$. Moreover, if ℓ is chosen maximal with the properties mentioned above, then we can also conclude that all μ_i with $\ell + 1 \leq i \leq r - 1$ contain an **active** position, which in turn means that $\beta_i \in \{\mathbf{o}\}^+$ for every i with $\ell + 1 \leq i \leq r - 1$. However, this directly implies that $\mu_i = \mathbf{a}$ with $\ell + 2 \leq i \leq r - 1$ and $\mu_{\ell+1} = \mathbf{c}^{g_2} \mathbf{a}$ for some $g_2 \geq 0$ with the property that at most one g_1 and g_2 can be positive. Consequently,

$$\widehat{\alpha}_1 = \mu_1 \underbrace{\mathbf{a}}_{\beta_1} \mu_2 \underbrace{\mathbf{a}}_{\beta_2} \cdots \mu_{\ell-1} \underbrace{\mathbf{a}}_{\beta_{\ell-1}} \mu_\ell \underbrace{\mathbf{a} \mathbf{o}^{g_1}}_{\beta_\ell} \underbrace{\mathbf{c}^{g_2} \mathbf{a}}_{\mu_{\ell+1}} \beta_{\ell+1} \underbrace{\mathbf{a}}_{\mu_{\ell+2}} \cdots \beta_{r-2} \underbrace{\mathbf{a}}_{\mu_{r-1}} \beta_{r-1},$$

where $\beta_i \in \{\mathbf{o}\}^+$ for every i with $\ell + 1 \leq i \leq r - 1$, $\mu_i \in \{\mathbf{c}\}^+$ for every i with $1 \leq i \leq \ell$, and $g_1, g_2 \geq 0$ with $0 \in \{g_1, g_2\}$.

If, on the other hand, no such ℓ , $1 \leq \ell < r$, exists, then all the **active** position t_i are in μ_i for every i with $1 \leq i \leq r - 1$. In particular, this means that $\beta_i \in \{\mathbf{o}\}^+$ for every i with $1 \leq i \leq r - 1$, which forces all μ_i , $2 \leq i \leq r - 1$, to start and end with an **active** position, while μ_1 must have a rightmost **active** position and a leftmost **active** position only if $\beta_0 \neq \varepsilon$. Thus,

$$\widehat{\alpha}_1 = \beta_0 \underbrace{\mathbf{c}^g \mathbf{a}}_{\mu_1} \beta_1 \underbrace{\mathbf{a}}_{\mu_2} \cdots \beta_{r-2} \underbrace{\mathbf{a}}_{\mu_{r-1}} \beta_{r-1},$$

where $\beta_i \in \{\mathbf{o}\}^+$ for every i with $1 \leq i \leq r - 1$, $g \geq 0$, and $g > 0$ implies $\beta_0 = \varepsilon$.

Note that all these observations have also obvious analogues for the right side $\widehat{\alpha}_2$. We now use these observations to prove the following claims regarding the structure of $\mu_r = \nu_1 \mathbf{a} \nu_2$:

1. If $\nu_1 \neq \varepsilon$, then we have

$$\widehat{\alpha}_1 = \mu_1 \mathbf{a} \mu_2 \mathbf{a} \cdots \mu_{r-1} \mathbf{a}.$$

Proof: If $\nu_1 \neq \varepsilon$, then $\nu_1[1] = \mathbf{c}$, which implies that $\beta_{r-1}[\lceil \beta_{r-1} \rceil] = \mathbf{a}$ and therefore $\beta_{r-1} = \mathbf{a}$. This means that for $\ell = r - 1$, we have the case described above, i. e., where t_ℓ is the rightmost **active** position that is not in μ_ℓ and, since $\beta_\ell = \mathbf{a}$, we also have the case $g_1 = 0$. This directly implies the statement claimed above. \square

2. If $\nu_2 \neq \varepsilon$, then we have

$$\widehat{\alpha}_2 = \mathbf{a} \mu_{r+1} \mathbf{a} \mu_{r+2} \cdots \mathbf{a} \mu_k.$$

Proof: Analogous to Claim 1. \square

3. If $\nu_1 = \varepsilon$, then we have one of the following two cases:

- (a) For some ℓ with $1 \leq \ell \leq r - 1$,

$$\widehat{\alpha}_1 = \mu_1 \underbrace{\mathbf{a}}_{\beta_1} \mu_2 \underbrace{\mathbf{a}}_{\beta_2} \cdots \mu_{\ell-1} \underbrace{\mathbf{a}}_{\beta_{\ell-1}} \mu_\ell \underbrace{\mathbf{a} \mathbf{o}^{g_1}}_{\beta_\ell} \underbrace{\mathbf{c}^{g_2} \mathbf{a}}_{\mu_{\ell+1}} \beta_{\ell+1} \underbrace{\mathbf{a}}_{\mu_{\ell+2}} \cdots \beta_{r-2} \underbrace{\mathbf{a}}_{\mu_{r-1}} \beta_{r-1},$$

where $\beta_i \in \{\mathbf{o}\}^+$ for every i with $\ell + 1 \leq i \leq r - 1$, $\mu_i \in \{\mathbf{c}\}^+$ for every i with $1 \leq i \leq \ell$, and $g_1, g_2 \geq 0$ with $0 \in \{g_1, g_2\}$.

- (b) $\widehat{\alpha}_1 = \beta_0 \underbrace{\mathbf{c}^g \mathbf{a}}_{\mu_1} \beta_1 \underbrace{\mathbf{a}}_{\mu_2} \cdots \beta_{r-2} \underbrace{\mathbf{a}}_{\mu_{r-1}} \beta_{r-1}$,

where $\beta_i \in \{\mathbf{o}\}^+$ for every i with $1 \leq i \leq r - 1$, $g \geq 0$, and $g > 0$ implies $\beta_0 = \varepsilon$.

Proof: If there is some ℓ' , $1 \leq \ell' \leq r - 1$, such that $t_{\ell'}$ is not in $\mu_{\ell'}$, then we can consider a maximal ℓ with this property and can conclude the statement of Claim 3a as observed above. If, on the other hand, no such ℓ' exists, then, as observed above, the statement of Claim 3b follows. \square

4. If $\nu_2 = \varepsilon$, then we have one of the following two cases:

(a) For some ℓ with $r \leq \ell \leq k$,

$$\widehat{\alpha}_2 = \beta_r \underbrace{\mathbf{a}}_{\mu_{r+1}} \beta_{r+1} \underbrace{\mathbf{a}}_{\mu_{r+2}} \cdots \beta_{\ell-1} \underbrace{\mathbf{a} \mathbf{c}^{g_1}}_{\mu_\ell} \underbrace{\mathbf{o}^{g_2} \mathbf{a}}_{\beta_\ell} \underbrace{\mu_{\ell+1}}_{\beta_{\ell+1}} \underbrace{\mathbf{a}}_{\mu_{\ell+2}} \underbrace{\mu_{\ell+2}}_{\beta_{\ell+2}} \cdots \underbrace{\mathbf{a}}_{\beta_{k-1}} \mu_k,$$

where $\beta_i \in \{\mathbf{o}\}^+$ for every i with $r \leq i \leq \ell-1$, $\mu_i \in \{\mathbf{c}\}^+$ for every i with $\ell+1 \leq i \leq k$, and $g_1, g_2 \geq 0$ with $0 \in \{g_1, g_2\}$.

(b) $\widehat{\alpha}_2 = \beta_r \underbrace{\mathbf{a}}_{\mu_{r+1}} \beta_{r+1} \underbrace{\mathbf{a}}_{\mu_{r+2}} \cdots \underbrace{\mathbf{a} \mathbf{c}^g}_{\mu_1} \beta_k$,

where $\beta_i \in \{\mathbf{o}\}^+$ for every i with $r \leq i \leq k-1$, $g \geq 0$, and $g > 0$ implies $\beta_k = \varepsilon$.

Proof: Analogous to Claim 3. □

We are now ready to conclude the proof. To this end, we have to consider the four cases depending on whether the factors ν_1 and ν_2 are empty or not. In the following cases, we also use the observation that occurrences of x_s must be contained in marked factors.

- $\nu_1 \neq \varepsilon$ and $\nu_2 \neq \varepsilon$: Claims 1 and 2 directly imply that $\alpha[j]$ is the only occurrence of x_s . Therefore, marking x_s has joined two marked blocks and did not change any other block. Hence, at step $i-1$ there were $k+1$ marked blocks, which is a contradiction.
- $\nu_1 = \varepsilon$ and $\nu_2 \neq \varepsilon$: With Claim 2, we know that there are no occurrences of x_s in the right side. Furthermore, Claim 3 means that we have the situation described in Claim 3a or the one of Claim 3b, which we shall treat as separate cases:
 - The statement of Claim 3a applies: If $\ell = r-1$, then there are no occurrences of x_s on the left side, which implies that $\alpha[j]$ is the only occurrence of x_s . Hence, since $\nu_2 \neq \varepsilon$, at step $i-1$ of σ there were k marked blocks, which is a contradiction. If, on the other hand, $\ell < r-1$, then every occurrence of x_s on the left side has at least one adjacent position that is **open**. Moreover, since $\beta_{r-1} \in \{\mathbf{o}\}^+$, also the occurrence $\alpha[j]$ has an adjacent position that is **open**. Consequently, all occurrences of x_s have at least one adjacent position that is **open**. Since the only **active** positions that can be **closed** in the next step of Q are occurrences of x_s , it is not possible to set an **active** position to **closed** in the next step, which means that an **open** position will be set to **active**. Hence, there are $k+1$ **active** positions in step p_s+1 of Q . This means that the first of the two cases from the beginning of the proof holds.
 - The statement of Claim 3b applies: All occurrences of x_s (including $\alpha[j]$) have at least one adjacent position that is **open**. It follows that there are $k+1$ **active** positions in step p_s+1 of Q , which means that the first of the two cases from the beginning of the proof holds.
- $\nu_1 \neq \varepsilon$ and $\nu_2 = \varepsilon$: This is analogous to the previous case.
- $\nu_1 = \nu_2 = \varepsilon$: If the statement of Claim 3a applies with $\ell < r-1$ or $g_1 > 0$, or the statement of Claim 3b applies, then all occurrences of x_s on the left side have an adjacent position that is **open**. Likewise, if the statement of Claim 4a applies with $\ell > r$ or $g_2 > 0$, or the statement of Claim 4b applies, then all occurrences of x_s on the right side have an adjacent position that is **open**. Furthermore, the occurrence $\alpha[j]$ also has an adjacent position that is **open**. Hence, there are $k+1$ **active** positions in step p_s+1 of Q and therefore the first of the two cases from the beginning of the proof holds.

If the statement of Claim 3a applies with $\ell = r-1$ and $g_1 = 0$, then there is no occurrence of x_s on the left side and the occurrence $\alpha[j]$ has an **active** position to its left. If now the statement of Claim 4a applies with $\ell > r$ or $g_2 > 0$, or the statement of Claim 4b applies, then the occurrence $\alpha[j]$ has an **open** position to its right, and also all occurrences of x_s on the right side have an adjacent position that is **open**. Consequently, there are $k+1$ **active** positions in step p_s+1 of Q and therefore the first of the two cases from the beginning of the proof holds. For the situation that the statement of Claim 4a applies with $\ell = r$

and $g_2 = 0$, but also the statement of Claim 3a applies with $\ell < r - 1$ or $g_1 > 0$, or the statement of Claim 3b applies, we can analogously conclude that there are $k + 1$ active positions in step $p_s + 1$ of Q .

The only remaining case is that the statement of Claim 3a applies with $\ell = r - 1$ and $g_1 = 0$ and the statement of Claim 4a applies with $\ell = r$ and $g_2 = 0$. We note that this implies the following:

$$\widehat{\alpha} = \mu_1 \underbrace{\mathbf{a}}_{\beta_1} \mu_2 \underbrace{\mathbf{a}}_{\beta_2} \cdots \mu_{r-1} \underbrace{\mathbf{a}}_{\beta_{r-1}} \underbrace{\mathbf{a}}_{\mu_r} \underbrace{\mathbf{a}}_{\beta_r} \mu_{r+1} \underbrace{\mathbf{a}}_{\beta_{r+1}} \cdots \mu_{k-1} \underbrace{\mathbf{a}}_{\beta_{k-1}} \mu_k.$$

This means that there is exactly one occurrence of x_s . In step $s - 1$ of the marking sequence σ there are exactly $k - 1$ marked blocks and, by our assumption that step s is the first step with k marked blocks, we also know that in steps $1, 2, \dots, s - 1$ the maximum number of marked blocks is $k - 1$. Moreover, at step $s - 1$, every unmarked position is adjacent to a marked block except the single occurrence of x_s that is marked in step s . Consequently, we can change σ into a marking sequence σ' as follows. The marking sequence σ' simulates σ up to step $s - 1$. So far, the maximum number of marked blocks is $k - 1$. Then, instead of marking x_s , σ' marks all other unmarked symbols in some order. Each of these marking steps leaves the number of marked blocks unchanged, or decreases it (this can be easily seen by consulting the factorisation illustrated above). Finally, symbol x_s is marked as the last symbol. Thus, σ' is a marking sequence for α with $\pi_{\sigma'}(\alpha) = k - 1$. This implies that the second of the two cases from the beginning of the proof holds. \square

Corollary 10. *Let α be a condensed word with $|\alpha| \geq 2$. Then $\text{loc}(\alpha) \leq \text{pw}(G_\alpha) \leq 2 \text{loc}(\alpha)$.*

Note that Corollary 10 is not true for condensed words α of size 1, since then $\text{loc}(\alpha) = 1$ and $\text{pw}(G_\alpha) = 0$. The reason why $\text{pw}(G_\alpha)$ can range between $\text{loc}(\alpha)$ and $2 \text{loc}(\alpha)$ (rather than $\text{pw}(G_\alpha) = 2 \text{loc}(\alpha)$) is that in a marking sequence, every marked block accounts for one unit of the quantity $\text{loc}(\alpha)$, while in the path decomposition, a marked block is represented either by two active vertices or by only one (if the block has size one). There are (condensed) examples that reach the extremes $\text{loc}(\alpha)$ and $2 \text{loc}(\alpha)$, i.e., the bounds of Corollary 10 are tight.

Proposition 11. *Let $\alpha = (x_1 x_2 \dots x_n x_{n-1} \dots x_2)^k x_1$ with $n \geq 3$, and let $\beta = (x_1 x_2)^k$. Then we have $\text{loc}(\alpha) = k$ and $\text{pw}(G_\alpha) = 2k$, and $\text{loc}(\beta) = \text{pw}(G_\beta) = k$.*

Proof. We start with proving the first statement and first observe that $\text{loc}(\alpha) \leq k$ due to the marking order x_n, x_{n-1}, \dots, x_1 . In order to show $\text{pw}(G_\alpha) \geq 2k$, we first observe that, for every $i \in \{2, \dots, n - 1\}$, $\text{ps}_{x_i}(\alpha)$ is a clique of size $2k$ in G_α , which implies that every path-decomposition for G_α reaches the state where all $2k$ vertices of $\text{ps}_{x_i}(\alpha)$ are active. Now let Q be a path-decomposition for G_α , let $i \in \{2, \dots, n - 1\}$ be such that all $\text{ps}_{x_i}(\alpha)$ are first set to active, i.e., when all vertices $\text{ps}_{x_i}(\alpha)$ are active for the first time, then in every $\text{ps}_{x_j}(\alpha)$, $j \in \{2, \dots, n - 1\} \setminus \{i\}$, there is at least one open vertex (in particular, no vertex from any ps_{x_j} , $2 \leq j \leq n - 1$, is closed). Moreover, in the following we consider the earliest point of Q , where all $\text{ps}_{x_i}(\alpha)$ are active.

If, at this point, there is some additional active vertex, then there are $2k + 1$ active vertices; thus, in the following we assume that there are no other active vertices. If there is also no closed vertex, then all other vertices are open, which means that every vertex from $\text{ps}_{x_i}(\alpha)$ has at least one adjacent open vertex and therefore we have to set an open vertex to active, before we can set a vertex from $\text{ps}_{x_i}(\alpha)$ to closed; this leads to at least $2k + 1$ active vertices. It remains to consider the case where there is some closed vertex j . This means that all vertices of $\text{ps}_{\alpha[j]}(\alpha)$ are closed, which implies that $j \in \text{ps}_{x_1}(\alpha) \cup \text{ps}_{x_n}(\alpha)$. We first consider the case $j \in \text{ps}_{x_1}(\alpha)$. Since every vertex from $\text{ps}_{x_2}(\alpha)$ is adjacent to some vertex from $\text{ps}_{x_1}(\alpha)$, we can conclude that all vertices from $\text{ps}_{x_2}(\alpha)$ are active, i.e., $i = 2$. The assumption $j \in \text{ps}_{x_n}(\alpha)$ analogously leads to the situation that $i = n - 1$. Consequently, all $2k$ vertices from $\text{ps}_{x_i}(\alpha)$ are active, either $\text{ps}_{x_1}(\alpha)$ are all closed or $\text{ps}_{x_n}(\alpha)$ are all closed, and all other vertices are open. In both of these cases, every vertex from $\text{ps}_{x_i}(\alpha)$ has at least one adjacent open vertex, which, as before, means that we have to set an open vertex to active, before we can set a vertex

from $\text{ps}_{x_i}(\alpha)$ to **closed**; this, as well, leads to at least $2k + 1$ **active** vertices. Consequently, $w(Q) \geq 2k$, and, with Corollary 10, we can conclude $\text{pw}(G_\alpha) = 2k$.

With respect to the second statement, we first note that any marking sequence for β leads to k marked blocks, which implies $\text{loc}(\beta) = k$. Moreover, a path-decomposition Q with $w(Q) = k$ can be easily constructed as follows. First, we set all positions of $\text{ps}_{x_1}(\beta)$ to **active**. Then we set position 2 to **active**, position 1 to **closed**, position 4 to **active**, position 3 to **closed** and so on, until all positions of $\text{ps}_{x_2}(\beta)$ are **active** and all positions of $\text{ps}_{x_1}(\beta)$ are **closed**. Finally, the positions of $\text{ps}_{x_2}(\beta)$ are set to **closed**. There are at most $k + 1$ positions **active** at the same time; thus, $w(Q) = k$ and therefore $\text{pw}(G_\beta) \leq k$. Together with Corollary 10, this implies $\text{loc}(\beta) = \text{pw}(G_\beta) = k$. \square

Note that the construction of a graph G_α from a word α does not technically provide a reduction from the decision problem **LOC** to **PATHWIDTH** (due to the fact that $\text{pw}(G_\alpha)$ lies between $\text{loc}(\alpha)$ and $2\text{loc}(\alpha)$) and therefore cannot be used to solve **MINLOC** exactly. Its main purpose is to carry over approximation results from **MINPATHWIDTH** to **MINLOC**, which is formally stated by the next lemma (in this regard, note that exact algorithms for **MINLOC** are obtained in Section 3 via a reduction to **MINCUTWIDTH** instead).

Lemma 12. *If there is an $r(\text{opt}, n)$ -approximation algorithm for **MINPATHWIDTH** running in $O(f(n))$ time, then there is an $2r(2\text{opt}, |\alpha|)$ -approximation algorithm for **MINLOC** with running time $O(f(|\alpha|) + |\alpha|^2)$.*

Proof. Let α be an instance of **MINLOC** and \mathcal{A} an $r(\text{opt}, n)$ -approximation for **MINPATHWIDTH**. By Corollary 10, it follows that $\text{pw}(G_\alpha) \leq 2m^*(\alpha)$.

In the proof of Corollary 10 it is shown that any path decomposition Q for G_α can be translated in time $O(|\alpha|)$ into a marking sequence σ with $\pi_\sigma(\alpha) \leq \text{pw}(Q)$. With the inequality $m^*(\alpha) \geq \frac{1}{2} \text{pw}(G_\alpha)$, the performance ratio of σ can be bounded by $R(\alpha, \sigma) = \frac{\pi_\sigma(\alpha)}{m^*(\alpha)} \leq \frac{2}{\text{pw}(G_\alpha)} \text{pw}(Q) \leq 2R(G_\alpha, Q)$. With $R(G_\alpha, Q) \leq r(\text{cw}(G_\alpha), n)$ from the approximation ratio of α , $n = |\alpha|$ from the construction of G_α , and $\text{cw}(G_\alpha) \leq 2m^*(\alpha)$ from Corollary 10, the claimed bound of $2r(2\text{opt}, |\alpha|)$ on the approximation ratio follows. The approximation procedure to compute σ , creates G_α in $O(|\alpha|^2)$, runs \mathcal{A} in $O(f(|\alpha|))$ and translates the path-decomposition Q into σ in $O(|\alpha|)$, which takes an overall running time in $O(f(|\alpha|) + |\alpha|^2)$. \square

Consequently, approximation algorithms for **MINPATHWIDTH** carry over to **MINLOC**. To the knowledge of the authors, the currently best approximation algorithm for **MINPATHWIDTH** is due to [20], with approximation ratio of $O(\sqrt{\log(\text{opt})} \log(n))$. This implies the following.

Theorem 13. *There is an $O(\sqrt{\log(\text{opt})} \log(n))$ -approximation algorithm for **MINLOC**.*

Another consequence that is worth mentioning is due to the fact that an optimal path decomposition can be computed faster than $O^*(2^n)$. More precisely, it is shown in [47] that for computing path decompositions, there is an exact algorithm with running time $O^*((1.9657)^n)$, and even an additive approximation algorithm with running time $O^*((1.89)^n)$. Consequently, there is a 2-approximation algorithm for **MINLOC** with running time $O^*((1.9657)^n)$ and an asymptotic 2-approximation algorithm with running time $O^*((1.89)^n)$ for **MINLOC**.

By combining the reduction from **MINCUTWIDTH** to **MINLOC** from Section 3 with the reduction from **MINLOC** to **MINPATHWIDTH** defined above, we obtain a reduction from **MINCUTWIDTH** to **MINPATHWIDTH** that carries over the pathwidth-approximation from [20] to **MINCUTWIDTH** as follows (in particular, this improves the state-of-the-art approximation for **MINCUTWIDTH** from [36]).

Theorem 14. *There is an $O(\sqrt{\log(\text{opt})} \log(n))$ -approximation algorithm for **MINCUTWIDTH**.*

Note that Theorem 14 only applies to **MINCUTWIDTH** for simple graphs; the case of multi-graphs shall be briefly discussed in Section 5.

Many existing algorithms constructing path decompositions are of theoretical interest only, and this disadvantage carries over to the possible algorithms computing the locality number or cutwidth based on them. However, the reduction of Corollary 10 is also applicable in a purely practical scenario, since any kind of practical algorithm constructing path decompositions can be

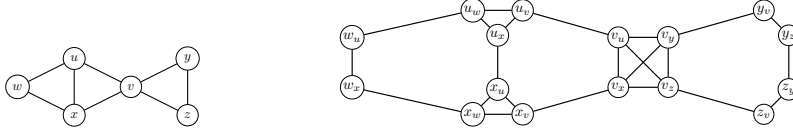


Figure 4: A graph G and the corresponding graph G' obtained by the reduction.

used in order to compute marking sequences (the additional tasks of building G_α and the translation of a path decomposition for it back to a marking sequence are computationally simple). This observation is particularly interesting since developing practical algorithms constructing tree and path decompositions of small width is a vibrant research area.²

5 Pathwidth and Cutwidth

Since pathwidth and cutwidth are classical graph parameters that play an important role for graph algorithms, independent from our application for computing the locality number, we also present a direct reduction from MINCUTWIDTH to MINPATHWIDTH.

For an arbitrary graph $G = (V, E)$, we construct the graph $G' = (V', E')$ with $V' = \{v_u \mid \{u, v\} \in E\}$ and $E' = \{\{u_v, v_u\} \mid \{u, v\} \in E\} \cup \{\{v_u, v_w\} \mid \{u, v\}, \{w, v\} \in E, u \neq w\}$. See Figure 4 for an example of the reduction.

For the next result, we first need the following definition. The *second order cutwidth* of a linear arrangement L is defined by $\text{cw}_2(L) = \max\{|\mathcal{C}_L(i-1) \cup \mathcal{C}_L(i)| \mid 1 \leq i \leq n+1\}$; the second order cutwidth of a graph is then defined by $\text{cw}_2(G) = \min\{\text{cw}_2(L) \mid L \text{ is a linear arrangement for } G\}$.

Lemma 15. *Let G be a graph with at least one edge. Then $\text{cw}(G) \leq \text{pw}(G') \leq 2 \text{cw}(G)$.*

Proof. Consider a graph $G = (V, E)$. For the inequality $\text{pw}(G') \leq 2 \text{cw}(G)$ let $L = (v_1, \dots, v_n)$ be an optimal linear arrangement for V . We show that the pathwidth of G' is at most $\text{cw}_2(G)$ which, by the definition of the second order cutwidth immediately yields $\text{pw}(G') \leq 2 \text{cw}(G)$.

To proof the claimed bound on the pathwidth of G' , we construct a path decomposition for G' of width at most $\text{cw}_2(G)$ as follows. The decomposition contains a bag for each edge $\{u, w\} \in E$, indexed by the ordered pair (i, j) if $u = v_i$ and $w = v_j$ with $i < j$ in L , and one bag for each vertex $u \in V$, indexed by $(i, 0)$ where $u = v_i$ in L . Formally, we construct the set of bags:

$$B = \{B(i, j) \mid \{v_i, v_j\} \in E, i < j\} \cup \{B(i, 0) \mid 1 \leq i \leq n\}.$$

We define the path decomposition P to be the bags in B listed in lexicographical order of the index-pairs.

Each bag $B(i, j)$ contains the union of the three vertex sets:

$$\begin{aligned} V'_{\geq}(i, j) &= \{u_w \in V' \mid u = v_i, w = v_\ell, \ell \geq j\} \\ V'_{\leq}(i, j) &= \{u_w \in V' \mid w = v_i, u = v_\ell, i < \ell \leq j\} \\ V'_{\ll}(i, j) &= \{u_w \in V' \mid u = v_x, w = v_y, y < i < x\} \end{aligned}$$

Each bag $B(i, j)$ has cardinality at most:

$$\begin{aligned} &|V'_{\geq}(i, j)| + |V'_{\leq}(i, j)| + |V'_{\ll}(i, j)| \\ &\leq |\{v_k \mid \{v_i, v_k\} \in E, k \geq j\}| + |\{v_k \mid \{v_i, v_k\} \in E, i < k \leq j\}| + |\{\{v_k, v_\ell\} \mid k < i < \ell\}| \\ &\leq |\{\{v_k, v_\ell\} \in E \mid k \leq i \leq \ell\}| + 1. \end{aligned}$$

This yields the claimed bound of $\text{cw}_2(G)$ on the pathwidth of G' .

It remains to prove that P is a feasible path decomposition for G' .

²See, e.g., the work [13] and the references therein for practical algorithms constructing path decompositions; also note that designing exact and heuristic algorithms for constructing tree decompositions was part of the ‘‘PACE 2017 Parameterized Algorithms and Computational Experiments Challenge’’ [16].

1. P satisfies the cover property:

For each edge $\{u_w, u_u\} \in E'$ let $u = v_i$ and $w = v_j$ in L with $i < j$, the bag $B(i, j)$ contains both u_w and u_u , as $u_w \in V'_\geq(i, j)$ and $u_u \in V'_\leq(i, j)$.

For each edge $\{v_u, v_w\} \in E'$ with $\{u, v\}, \{w, v\} \in E, u \neq w$, let $v = v_i$, then the bag $B(i, 0)$ contains both v_u and v_w , since which ever index u and w have, it is larger than 0, so $v_u, v_w \in V'_\geq(i, 0)$.

These are the only two types of edges in G' , so P satisfies the cover property.

2. P satisfies the connectivity property:

We consider a vertex $u_w \in V'$ and show that the bags containing u_w are a consecutive set in P . To this end we distinguish for $u = v_x$ and $w = v_y$ in L , the two cases $x < y$ and $x > y$.

If $x < y$, we claim that the set of bags in P which contain u_w is exactly the consecutive set $S = \{B(x, r) \mid 0 \leq r \leq y, \{v_x, v_r\} \in E \text{ or } r = 0\}$.

For each $0 \leq r \leq y$, u_w lies in $V'_\geq(x, r) \subseteq B(x, r)$.

For each bag $B(i, j)$ in $B \setminus S$, case analysis shows that $u_w \notin B(i, j)$:

- (i, j) smaller than $(x, 0)$ in the order on P , so $i < x < y$:
 - $u = v_x \neq v_i$, so $u_w \notin V'_\geq(i, j)$.
 - $w = v_y \neq v_i$, so $u_w \notin V'_\leq(i, j)$.
 - $u = v_x$ with $i < x$, so $u_w \notin V'_{\leq}(i, j)$.
- (i, j) larger than (x, y) in the order on P with $i = x$, so $y < j$:
 - $u = v_i$ but $w = v_y$ with $y < j$, so $u_w \notin V'_\geq(i, j)$.
 - $w = v_y \neq v_i$, so $u_w \notin V'_\leq(i, j)$.
 - $u = v_x$ with $x = i$, so $u_w \notin V'_{\leq}(i, j)$.
- (i, j) larger than (x, y) in the order on P with $i > x$:
 - $u \neq v_i$, so $u_w \notin V'_\geq(i, j)$.
 - $u = v_x$ with $x < i$, so $u_w \notin V'_\leq(i, j)$.
 - $u = v_x$ with $x < i$, so $u_w \notin V'_{\leq}(i, j)$.

If $x > y$, we claim that the set of bags in P which contain u_w is exactly the consecutive set: $S = \{B(y, x), \dots, B(x, 0)\} = \{B(i, j) \mid (y, x) \leq (i, j) \leq (x, 0), \{v_i, v_j\} \in E \text{ or } j = 0\}$.

Each bag in S contains u_w :

- For each $y < i < x$, and j with $\{v_i, v_j\} \in E$ or $j = 0$, u_w lies in $V'_\geq(i, j) \subseteq B(i, j)$.
- For each j with $x \leq j$, u_w lies in $V'_\leq(y, j) \subseteq B(y, j)$.
- Since $y \geq 0$, $u_w \in V'_\geq(x, 0) \subseteq B(x, 0)$.

For each bag $B(i, j)$ in $B \setminus S$, case analysis shows that $u_w \notin B(i, j)$:

- (i, j) smaller than (y, x) in the order on P with $i = y$, so $j < x$:
 - $u \neq v_i$, so $u_w \notin V'_\geq(i, j)$
 - $w = v_i$, but $u = v_x$ with $x > j$, so $u_w \notin V'_\leq(i, j)$.
 - $w = v_y$ with $i = y$, so $u_w \notin V'_{\leq}(i, j)$.
- (i, j) smaller than (y, x) in the order on P with $i < y$, so $i < y < x$:
 - $u \neq v_i$ so $u_w \notin V'_\geq(i, j)$.
 - $w \neq v_i$ so $u_w \notin V'_\leq(i, j)$.
 - $w = v_y$ with $i < y$, so $u_w \notin V'_{\leq}(i, j)$.
- (i, j) larger than $(x, 0)$ in the order on P with $i = x$, so $j > i$ (observe that we only define bags $B(i, j)$ with $i < j$ or $j = 0$):

- $u = v_i$ but $w = v_y$ with $y < x = i \leq j$, so $u_w \notin V'_{\geq}(i, j)$.
 - $w = v_y \neq v_i$, so $u_w \notin V'_{\leq}(i, j)$.
 - $u = v_x$ with $x = i$, so $u_w \notin V'_{\leq}(i, j)$.
- (i, j) larger than $(x, 0)$ in the order on P with $i > x$, so $y < x < i$:
 - $u \neq v_i$, so $u_w \notin V'_{\geq}(i, j)$.
 - $v \neq v_i$, so $u_w \notin V'_{\leq}(i, j)$.
 - $x < i$, so $u_w \notin V'_{\leq}(i, j)$.

This concludes the proof for the inequality $\text{pw}(G') \leq 2\text{cw}(G)$.

To show the other inequality $\text{cw}(G) \leq \text{pw}(G')$, we will actually show the even stronger result: Given a path decomposition of width k for G' , it is possible to construct a linear arrangement with cutwidth at most k for G in polynomial time.

Let $P = \{B_1, \dots, B_r\}$ be a path decomposition of width k for G' . Since the vertices in $N_v := \{v_u \mid \{u, v\} \in E\}$ form a clique in G' , there has to be at least one bag in P which contains N_v , for each $v \in V$. Pick for each $v \in V$ an index $\phi(v)$ with $N_v \subset B_{\phi(v)}$. Define the linear order $L = (v_1, \dots, v_n)$ on V according to the order on the indices $\phi(v)$. Let $t \in \{1, \dots, n\}$ be such that $|\mathcal{C}_L(t)| = \text{cw}(L)$. Recall the definition $\mathcal{C}_L(t) = \{\{v_i, v_j\} \mid i \leq t < j\}$. We will show that $\{u_w, w_u\} \cap B_{\phi(v_t)} \neq \emptyset$ for each edge $\{u, w\} \in \mathcal{C}_L(t)$.

For every pair of indices (i, j) with $i \leq t < j$ and $\{v_i, v_j\} \in E$, denote $u = v_i$ and $w = v_j$. By definition of G' , $\{u_w, w_u\} \in E'$, so by cover property there has to be at least one bag $B_x \in P$ which contains both u_w and w_u . The bag $B_{\phi(u)}$ contains u_w and the bag $B_{\phi(w)}$ contains w_u . By connectivity property, w_u has to be included in all bags between B_x and $B_{\phi(w)}$, and u_w has to be included in all bags between B_x with $B_{\phi(u)}$. Since $\phi(u) \leq \phi(v_t) < \phi(w)$, this implies that if $x \leq \phi(t)$, it follows that $w_u \in B_{\phi(v_t)}$ and otherwise if $\phi(t) \geq x$ it follows that $u_w \in B_{\phi(v_t)}$.

In fact, the above consideration also holds for (i, j) with $i \leq t \leq j$ and $\{v_i, v_j\} \in E$, so it follows that $|B_{\phi(v_t)}| \geq |\mathcal{C}_L(t-1) \cup \mathcal{C}_L(t)|$.

So if there is an index t' with $|\mathcal{C}_L(t')| = \text{cw}(L)$ and $|\mathcal{C}_L(t')| < |\mathcal{C}_L(t'-1) \cup \mathcal{C}_L(t')|$, then it follows that:

$$k \geq |B_{\phi(v_{t'})}| - 1 \geq |\mathcal{C}_L(t'-1) \cup \mathcal{C}_L(t')| - 1 \geq |\mathcal{C}_L(t')| = \text{cw}(L).$$

More generally, if there exists an index $j \in \{1, \dots, n\}$ such that $|B_{\phi(j)}| \geq \text{cw}(L) + 1$, the claimed bound of k on the value of L follows. We claim that if no such index j exists, we can construct a better linear arrangement in polynomial time, by rearranging all v_t with index $t \in I_{\max} = \{t \mid |\mathcal{C}_L(t)| = \text{cw}(L)\}$. For each such index $t \in I_{\max}$, we know that $\mathcal{C}_L(t) = \mathcal{C}_L(t-1) \cup \mathcal{C}_L(t)$ hence v_t has no neighbour in $\{v_1, \dots, v_{t-1}\}$. Move v_t to the right of its neighbour v_ℓ of smallest index in L . This way, the size of the cut for v_t is equal to the previous cut for v_ℓ which has to be smaller than $\text{cw}(L)$, since v_ℓ had v_t as a neighbour to the left which means that $\mathcal{C}_L(\ell) = \mathcal{C}_L(\ell-1) \cup \mathcal{C}_L(\ell)$ did not hold and ℓ could hence not have been in I_{\max} . The rearrangement of v_t can only increase the cut for v_ℓ . This can only happen, if v_t has degree 1, as otherwise the cut for v_ℓ does no longer count at least one edge adjacent to v_t which makes up for the edge $\{v_t, v_\ell\}$ which is added by the rearrangement. Assuming a connected input graph with at least three vertices, this implies that v_ℓ has at least one neighbour other than v_t . If the cut of v_ℓ increases to $\text{cw}(L)$ by moving v_t to the right of v_ℓ , then the cut value of v_ℓ was $\text{cw}(L) - 1$ in the unaltered arrangement while $|B_{\phi(\ell)}| \leq \text{cw}(L)$, so v_t was the only neighbour of v_ℓ with index smaller than ℓ . In this case, move v_t directly to the left of v_ℓ , then the cut of v_ℓ remains $\text{cw}(L) - 1$ and the cut of v_t is computed from the edges crossing both v_ℓ and v_t plus 1 for the edge $\{v_\ell, v_t\}$ minus at least 1 for the edges from v_ℓ to a neighbour in $\{v_{\ell+1}, \dots, v_n\}$, hence also at most $\text{cw}(L) - 1$. Repeating this procedure for each $t \in I_{\max}$ yields a linear arrangement with cut at most $\text{cw}(L) - 1 \leq |B_{\phi(v_t)}| - 1 \leq k$. \square

Lemma 15 does not only prove that $\text{cw}(G) \leq \text{pw}(G') \leq 2\text{cw}(G)$, but also yields a constructive way to compute a linear arrangement for G of cut at most k from a path decomposition of width k for G' . Further, Lemma 15 remains true if G is a multigraph; observe that the reduction still constructs a simple graph G' . This gives the following result.

Lemma 16. *If there is an $r(\text{opt}, |V|)$ -approximation algorithm for MINPATHWIDTH with running-time $O(f(|V|))$, then there is an $2r(2 \text{opt}, h)$ -approximation algorithm for MINCUTWIDTH on multigraphs with running time $O(f(h) + h^2 + n)$, where n is the number of vertices and h is the number of edges.*

Proof. Let $G = (V, E)$ be an instance of MINCUTWIDTH with and \mathcal{A} an $r(\text{pw}(G'), |V|)$ -approximation for MINPATHWIDTH. By Lemma 15, it follows that $\text{cw}(G) \geq \frac{1}{2} \text{pw}(G')$.

Further, the proof of Lemma 15 shows that a path-decomposition P of width k for G' can be translated into a linear arrangement L for G with $\text{cw}(L) \leq k$ in $O(h^2 + n)$. The relative error of L can hence be bounded by $R(G, L) = \frac{\text{cw}(L)}{\text{cw}(G)} \leq \frac{2 \text{pw}(P)}{\text{pw}(G')} = 2R(G', P)$. The algorithm which builds G' from G in $O(n + h)$, runs \mathcal{A} on G' in $O(f(h))$ and creates a linear arrangement L in $O(h^2 + n)$ has a performance ratio $2r(\text{pw}(G'), |V|) \leq 2r(2 \text{cw}(G), h)$ and an overall running time in $O(f(h) + h^2 + n)$. \square

With the $O(\sqrt{\log(\text{opt})} \log(n))$ -approximation for MINPATHWIDTH from [20], Lemma 16 gives the following approximation for MINCUTWIDTH on multigraphs.

Theorem 17. *There is an $O(\sqrt{\log(\text{opt})} \log(h))$ -approximation algorithm for MINCUTWIDTH on multigraphs with h edges.*

In accordance with Thm. 13, Thm. 17 yields for simple graphs an $O(\sqrt{\log(\text{opt})} \log(n))$ -approximation algorithm. Analogously, Thm. 13 could be formulated for multigraphs, which would also change the approximation-ratio to $O(\sqrt{\log(\text{opt})} \log(h))$.

References

- [1] Amihod Amir and Igor Nor. Generalized function matching. *Journal of Discrete Algorithms*, 5(3):514–523, 2007. doi:10.1016/j.jda.2006.10.001.
- [2] Dana Angluin. Finding patterns common to a set of strings. *Journal of Computer and System Sciences*, 21(1):46–62, 1980. doi:10.1016/0022-0000(80)90041-0.
- [3] Sanjeev Arora, Boaz Barak, and David Steurer. Subexponential algorithms for unique games and related problems. *Journal of the ACM*, 62(5):42:1–42:25, 2015. doi:10.1145/2775105.
- [4] Sanjeev Arora, Satish Rao, and Umesh V. Vazirani. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM*, 56(2):5:1–5:37, 2009. doi:10.1145/1502793.1502794.
- [5] Giorgio Ausiello, Alberto Marchetti-Spaccamela, Pierluigi Crescenzi, Giorgio Gambosi, Marco Protasi, and Viggo Kann. *Complexity and approximation: combinatorial optimization problems and their approximability properties*. Springer, 1999. doi:10.1007/978-3-642-58412-1.
- [6] Boaz Barak, Prasad Raghavendra, and David Steurer. Rounding semidefinite programming hierarchies via global correlation. In Rafail Ostrovsky, editor, *52nd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2011*, pages 472–481. IEEE, 2011. doi:10.1109/focs.2011.95.
- [7] Pablo Barceló, Leonid Libkin, Anthony W. Lin, and Peter T. Wood. Expressive languages for path queries over graph-structured data. *ACM Transactions on Database Systems*, 37(4):1–46, 2012. doi:10.1145/2389241.2389250.
- [8] Hans L. Bodlaender. A tourist guide through treewidth. *Acta Cybernetica*, 11(1-2):1–21, 1993. URL: http://www.inf.u-szeged.hu/actacybernetica/edb/vol11n1_2/pdf/Bodlaender_1993_ActaCy
- [9] Hans L. Bodlaender. A linear-time algorithm for finding tree-decompositions of small treewidth. *SIAM Journal on Computing*, 25(5):1305–1317, 1996. doi:10.1137/s0097539793251219.

- [10] Hans L. Bodlaender. A partial k -arboretum of graphs with bounded treewidth. *Theoretical Computer Science*, 209(1–2):1–45, 1998. doi:10.1016/S0304-3975(97)00228-4.
- [11] Hans L. Bodlaender. Fixed-parameter tractability of treewidth and pathwidth. In Hans L. Bodlaender, Rod Downey, Fedor V. Fomin, and Dániel Marx, editors, *The Multivariate Algorithmic Revolution and Beyond*, volume 7370 of *LNCS*, pages 196–227, 2012. doi:10.1007/978-3-642-30891-8_12.
- [12] Hans L. Bodlaender, Fedor V. Fomin, Arie M. C. A. Koster, Dieter Kratsch, and Dimitrios M. Thilikos. A note on exact algorithms for vertex ordering problems on graphs. *Theory of Computing Systems*, 50(3):420–432, 2012. doi:10.1007/s00224-011-9312-0.
- [13] David Coudert, Dorian Mazauric, and Nicolas Nisse. Experimental evaluation of a branch-and-bound algorithm for computing pathwidth and directed pathwidth. *ACM Journal of Experimental Algorithmics*, 21(1):1.3:1–1.3:23, 2016. doi:10.1145/2851494.
- [14] Joel D. Day, Pamela Fleischmann, Florin Manea, and Dirk Nowotka. Local patterns. In Satya V. Lokam and R. Ramanujam, editors, *Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2017*, volume 93 of *LIPICs*, pages 24:1–24:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017. doi:10.4230/LIPICs.FSTTCS.2017.24.
- [15] Joel D. Day, Pamela Fleischmann, Florin Manea, Dirk Nowotka, and Markus L. Schmid. On matching generalised repetitive patterns. In Mizuho Hoshi and Shinnosuke Seki, editors, *Developments in Language Theory, DLT 2018*, volume 11088 of *LNCS*, pages 269–281. Springer, 2018. doi:10.1007/978-3-319-98654-8_22.
- [16] Holger Dell, Christian Komusiewicz, Nimrod Talmon, and Mathias Weller. The PACE 2017 parameterized algorithms and computational experiments challenge: The second iteration. In Daniel Lokshtanov and Naomi Nishimura, editors, *Parameterized and Exact Computation, IPEC 2017*, volume 89 of *LIPICs*, pages 30:1–30:12. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017. doi:10.4230/LIPICs.IPEC.2017.30.
- [17] Josep Díaz, Jordi Petit, and Maria Serna. A survey of graph layout problems. *ACM Computing Surveys*, 34(3):313–356, 2002. doi:10.1145/568522.568523.
- [18] Rodney G. Downey and Michael R. Fellows. *Fundamentals of Parameterized Complexity*. Springer, 2013. doi:10.1007/978-1-4471-5559-1.
- [19] Thomas Erlebach, Peter Rossmanith, Hans Stadtherr, Angelika Steger, and Thomas Zeugmann. Learning one-variable pattern languages very efficiently on average, in parallel, and by asking queries. *Theoretical Computer Science*, 261(1):119–156, 2001. doi:10.1016/S0304-3975(00)00136-5.
- [20] Uriel Feige, MohammadTaghi HajiAghayi, and James R. Lee. Improved approximation algorithms for minimum weight vertex separators. *SIAM Journal on Computing*, 38(2):629–657, 2008. doi:10.1137/05064299x.
- [21] Henning Fernau, Florin Manea, Robert Mercas, and Markus L. Schmid. Pattern matching with variables: Fast algorithms and new hardness results. In Ernst W. Mayr and Nicolas Ollinger, editors, *Symposium on Theoretical Aspects of Computer Science, STACS 2015*, volume 30 of *LIPICs*, pages 302–315. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2015. doi:10.4230/LIPICs.STACS.2015.302.
- [22] Henning Fernau, Florin Manea, Robert Mercas, and Markus L. Schmid. Revisiting Shinohara’s algorithm for computing descriptive patterns. *Theoretical Computer Science*, 733:44–54, 2018. doi:10.1016/j.tcs.2018.04.035.
- [23] Henning Fernau and Markus L. Schmid. Pattern matching with variables: A multivariate complexity analysis. *Information and Computation*, 242:287–305, 2015. doi:10.1016/j.ic.2015.03.006.

- [24] Henning Fernau, Markus L. Schmid, and Yngve Villanger. On the parameterised complexity of string morphism problems. *Theory of Computing Systems*, 59(1):24–51, 2016. doi:10.1007/s00224-015-9635-3.
- [25] Jörg Flum and Martin Grohe. *Parameterized Complexity Theory*. Springer, 2006. doi:10.1007/3-540-29953-X.
- [26] Dominik D. Freydenberger. Extended regular expressions: Succinctness and decidability. *Theory of Computing Systems*, 53(2):159–193, 2013. doi:10.1007/s00224-012-9389-0.
- [27] Dominik D. Freydenberger and Markus L. Schmid. Deterministic regular expressions with back-references. In Heribert Vollmer and Brigitte Vallée, editors, *Symposium on Theoretical Aspects of Computer Science, STACS 2017*, volume 66 of *LIPICs*, pages 33:1–33:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017. doi:10.4230/LIPICs.STACS.2017.33.
- [28] Jeffrey E. F. Friedl. *Mastering Regular Expressions*. O’Reilly, Sebastopol, CA, 3rd edition, 2006.
- [29] Venkatesan Guruswami and Ali Kemal Sinop. Lasserre hierarchy, higher eigenvalues, and approximation schemes for graph partitioning and quadratic integer programming with PSD objectives. In Rafail Ostrovsky, editor, *52nd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2011*, pages 482–491. IEEE, 2011. doi:10.1109/FOCS.2011.36.
- [30] Carl Hierholzer and Christian Wiener. ber die Möglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechung zu umfahren. *Mathematische Annalen*, 6(1):30–32, 1873. doi:10.1007/bf01442866.
- [31] Juhani Karhumki, Filippo Mignosi, and Wojciech Plandowski. The expressibility of languages and relations by word equations. *Journal of the ACM*, 47(3):483–505, 2000. doi:10.1145/337244.337255.
- [32] Michael Kearns and Leonard Pitt. A polynomial-time algorithm for learning k -variable pattern languages from examples. In Ronald L. Rivest, David Haussler, and Manfred K. Warmuth, editors, *Computational Learning Theory, COLT 1989*, pages 57–71. Morgan Kaufmann, 1989. doi:10.1016/b978-0-08-094829-4.50007-6.
- [33] Subhash Khot. On the power of unique 2-prover 1-round games. In John H. Reif, editor, *34th Annual ACM Symposium on Theory of Computing, STOC 2002*, pages 767–775. ACM, 2002. doi:10.1145/509907.510017.
- [34] Subhash Khot. On the unique games conjecture (invited survey). In *Computational Complexity, CCC 2010*, pages 99–121. IEEE, 2010. doi:10.1109/CCC.2010.19.
- [35] Ton Kloks, editor. *Treewidth, Computations and Approximations*, volume 842 of *LNCS*. Springer, 1994. doi:10.1007/BFb0045375.
- [36] Tom Leighton and Satish Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM*, 46(6):787–832, 1999. doi:10.1145/331524.331526.
- [37] M. Lothaire, editor. *Algebraic Combinatorics on Words*. Cambridge University Press, 2002. doi:10.1017/cbo9781107326019.
- [38] Fillia Makedon, Christos H. Papadimitriou, and Ivan Hal Sudborough. Topological bandwidth. *SIAM Journal on Algebraic and Discrete Methods*, 6(3):418–444, 1985. doi:10.1137/0606044.
- [39] Yen Kaow Ng and Takeshi Shinohara. Developments from enquiries into the learnability of the pattern languages from positive data. *Theoretical Computer Science*, 397(1–3):150–165, 2008. doi:10.1016/j.tcs.2008.02.028.
- [40] Jordi Petit. Addenda to the survey of layout problems. *Bulletin of the EATCS*, 105:177–201, 2011. URL: <http://eatcs.org/beatcs/index.php/beatcs/article/view/98>.

- [41] Prasad Raghavendra and David Steurer. Graph expansion and the unique games conjecture. In Leonard J. Schulman, editor, *42nd ACM Symposium on Theory of Computing, STOC 2010*, pages 755–764. ACM, 2010. doi:10.1145/1806689.1806792.
- [42] Prasad Raghavendra, David Steurer, and Madhur Tulsiani. Reductions between expansion problems. In *Computational Complexity, CCC 2012*, pages 64–73. IEEE, 2012. doi:10.1109/CCC.2012.43.
- [43] Daniel Reidenbach. Discontinuities in pattern inference. *Theoretical Computer Science*, 397(1–3):166–193, 2008. doi:10.1016/j.tcs.2008.02.029.
- [44] Daniel Reidenbach and Markus L. Schmid. Patterns with bounded treewidth. *Information and Computation*, 239:87–99, 2014. doi:10.1016/j.ic.2014.08.010.
- [45] Markus L. Schmid. Characterising REGEX languages by regular languages equipped with factor-referencing. *Information and Computation*, 249:1–17, 2016. doi:10.1016/j.ic.2016.02.003.
- [46] Takeshi Shinohara. Polynomial time inference of pattern languages and its application. In *7th IBM Symposium on Mathematical Foundations of Computer Science*, pages 191–209, 1982.
- [47] Karol Suchan and Yngve Villanger. Computing pathwidth faster than 2^n . In Jianer Chen and Fedor V. Fomin, editors, *Parameterized and Exact Computation, IWPEC 2009*, volume 5917 of *LNCS*, pages 324–335. Springer, 2009. doi:10.1007/978-3-642-11269-0_27.
- [48] Dimitrios M. Thilikos, Maria J. Serna, and Hans L. Bodlaender. Cutwidth I: A linear time fixed parameter algorithm. *Journal of Algorithms*, 56(1):1–24, 2005. doi:10.1016/j.jalgor.2004.12.001.
- [49] Yu (Ledell) Wu, Per Austrin, Toniann Pitassi, and David Liu. Inapproximability of treewidth and related problems. *Journal of Artificial Intelligence Research*, 49(1):569–600, 2014. doi:10.1613/jair.4030.

A Additional Word-Combinatorial Considerations

The Locality of the Zimin words

Proposition 18. $\text{loc}(Z_i) = \frac{|Z_i|+1}{4} = 2^{i-2}$ for $i \in \mathbb{N}_{\geq 2}$.

Proof. Clearly, x_1 and $x_1x_2x_1$ are 1-local. Consider a fixed $i \in \mathbb{N}$ and the marking sequence $(x_2, x_1, y_1, y_2, \dots, y_{i-2})$ for $i \geq 3$ and $\{y_1, \dots, y_{i-2}\} = \{x_3, \dots, x_i\}$. Notice that for all $j \in \mathbb{N}$, x_j occurs 2^{i-j} times in Z_i . Thus by marking x_2 , there are 2^{i-2} marked blocks. Since all occurrences of x_1 are adjacent to occurrences of x_2 , marking x_1 does not change the number of marked blocks. As marking the remaining variables only leads to the merging of some pairs of consecutive blocks into one, we never have more than 2^{i-2} marked blocks.

In the following we will show the converse. More precisely, we show that if a sequence is optimal for Z_i then it starts with x_2, x_1 . Let us note first that, for $2 \leq p < r$, between two consecutive occurrences of x_r in Z_i there is one occurrence of x_p . More precisely, each occurrence of a variable x_p , with $p \geq 2$, is directly between two occurrences of x_1 . Also, notice that x_j has 2^{i-j} occurrences in Z_i . Now, if x_1 is marked before x_2 , because Z_i starts with x_1x_2 and ends with x_2x_1 , it is immediate that after the marking of x_1 we will have at least $2^{i-2} + 1$ marked blocks in the word (separated by the 2^{i-2} unmarked occurrences of x_2). This is, thus, a marking sequence that is not optimal. So x_2 is marked before x_1 in an optimal sequence. Assume that there exists x_j , with $j > 2$, which is also marked before x_1 in an optimal sequence. Let w be a word such that $Z_i = x_1wx_1$. There are $2^{i-1} - 2$ occurrences of x_1 in w , and w starts with x_2x_1 and ends with x_1x_2 . As each two consecutive (marked) occurrences of the letters x_2 and x_j are separated by unmarked occurrences of x_1 we have that, just before marking x_1 , there are at least $\min\{2^{i-1} - 1, 2^{i-2} + 2^{i-j}\}$ marked blocks in w (and the same number in Z_i). This again shows that this is not an optimal marking sequence. So, before x_1 is marked, only x_2 should be marked. This concludes the proof of our claim, and of the proposition. \square

The Locality of (Condensed) Palindromes and Repetitions. We use the following notation. Given a marking sequence σ , let σ^R be the marking sequence obtained by reversing σ (i.e. $\sigma^R(i) = \sigma(|X| - i + 1)$ for $1 \leq i \leq |X|$).

By $\text{loc}(\text{cond}(w)) = \text{loc}(w)$, it is enough to show our results for condensed words. Since there are no condensed palindromes of even length, only palindromes of odd length are of interest when determining the locality number. A word w is called strictly k -local if for every optimal marking sequence of w there is a stage when exactly k factors are marked. For a letter $a \in \text{alph}(w)$, we denote by $|w|_a$ the number of occurrences of a in w . For simplicity of notations, let $[n] := \{1, 2, \dots, n\}$.

Let $w_i \in (X \cup \overline{X})^*$ be the marked version of w at stage $i \in [| \text{alph}(w) |]$ for a given marking sequence σ .

Lemma 19. Define the morphism $f : X \cup \overline{X} \rightarrow \{0, 1\}$ by

$$f(x) = \begin{cases} 0 & \text{if } x \in X, \\ 1 & \text{if } x \in \overline{X}. \end{cases}$$

If w is a palindrome and σ a marking sequence for w then $f(w_i)$ is a palindrome for all $i \in [| \text{alph}(w) |]$.

Proof. Let $w = uxu^R$ be a palindrome with $u \in X^*$ and $x \in X \cup \overline{X}$ and $|w| = n \in \mathbb{N}$. Moreover let σ be a marking sequence for w and $i \in [| \text{alph}(w) |]$. Since w is a palindrome, $w[j] = w[n - j]$. This implies $w_i[j], w_i[n - j]$ are both either in X or in \overline{X} . Thus either are both mapped to 0 or to 1. Consequently $f(w_i)$ is a palindrome. \square

Recall the definition of *border priority markable* from [14]. A strictly k -local word $w = avb \in XX^*X$ is called border priority markable if there exists a marking sequence σ of w such that in every stage $i \in [| \alpha(w) |]$ of σ where k blocks are marked, a and b are marked as well. Analogously right-border priority markable and left-border priority markable are defined: A strictly k -local word $w = avb \in XX^*X$ is called right-border priority markable (rbpm) if there exists a

marking sequence σ of w such that in every stage $i \in [|\alpha(w)|]$ of σ where k blocks are marked, b is marked as well - respectively, for left-border priority markable, a is marked as well.

Remark 20. If $w \in X^*$ is right-border priority markable, then u^R is left-border priority markable.

Lemma 21. *Let $w = uau^R$ be an odd-length condensed palindrome with $u \in X^*$ and $\mathbf{a} \in X$. Let u be strictly k -local witnessed by the marking sequence σ .*

- If u is rbpm then $\text{loc}(w) = 2k - 1$,
- if u is not rbpm and $\mathbf{a} \notin \text{alph}(u)$ then $\text{loc}(w) = 2k$,
- if u is not rbpm and $\mathbf{a} \in \text{alph}(u)$ and for all optimal marking sequences for u there exists a stage $i \in [|\text{alph}(u)|]$ such that \mathbf{a} is marked, k blocks are marked, and $u[[u]]$ is unmarked then $\text{loc}(w) = 2k + 1$, and
- else $\text{loc}(w) = 2k$.

Proof. Let σ be an optimal marking sequence of u . If $\mathbf{a} \in \text{alph}(u)$ then σ is a marking sequence for w . Marking w w.r.t. σ leads to $\pi_\sigma(w) \leq 2k + 1$ since there are at most maximal k blocks marked each in u and u^R , and additionally the single \mathbf{a} in the middle. If $\mathbf{a} \notin \text{alph}(u)$ then $\sigma' = \sigma \cup \{(|u| + 1, \mathbf{a})\}$ is a marking sequence for w with $\pi_{\sigma'}(w) \leq 2k$, since by marking w.r.t. σ maximal k blocks are marked by σ each in u and u^R and afterwards on marking a two blocks are joined. Thus in any case $\text{loc}(w) \leq 2k + 1$.

case 1. Consider u to be rbpm. Thus in every stage $i \in [|\text{alph}(u)|]$ where k blocks are marked, $u[[u]]$ is marked. This implies that $\pi_\sigma(w) \leq 2k - 1$ or $\pi_{\sigma'}(w) \leq 2k - 1$ with σ' defined as above.

Supposition: $\text{loc}(w) =: \ell < 2k - 1$

Let μ be an optimal marking sequence for w . Then μ is also a marking sequence for u and thus $\pi_\mu(u) \geq k$. By $\text{loc}(u) = k$ there exists a stage $i \in [|\text{alph}(w)|]$ of μ such that k blocks are marked in u , or more precisely $|\text{cond}(f(u_i))|_1 = k$. On the other hand $|\text{cond}(f(w_i))|_1 \leq \ell$. Since u is rbpm $u[[u]]$ is marked. If x is not marked, $|\text{cond}(f(u_i))|_1 \leq \frac{\ell}{2} < \frac{2k-1}{2} = k - \frac{1}{2}$. If x is marked, $|\text{cond}(f(u_i))|_1 \leq \frac{\ell-1}{2} < \frac{2k-2}{2} = k - 1$. This is in both cases a contradiction to $|\text{cond}(f(u_i))|_1 = k$.

case 2. Consider now that u is not rbpm. Thus there exists a stage $i \in [|\text{alph}(u)|]$ in which k blocks are marked but $u[[u]]$ is unmarked. If \mathbf{a} is not in $\text{alph}(u)$ marking \mathbf{a} before stage i leads to $2k + 1$ blocks for the largest such i . Considering σ' then at the beginning u and u^R are completely marked and in the end two blocks are joined by marking \mathbf{a} . This leads to $\text{loc}(w) \leq 2k$.

Supposition: $\text{loc}(w) < 2k$

As described, \mathbf{a} needs to be marked after the last stage where in u k blocks are marked without $u[[u]]$ being marked. But this sums up to k blocks marked in u and k blocks marked in u^R , hence overall $2k$ blocks. This concludes the case $\mathbf{a} \notin \text{alph}(u)$.

Consider $\mathbf{a} \in \text{alph}(u)$ and assume that \mathbf{a} is marked by σ when k blocks are marked in u and $u[[u]]$ is unmarked. Thus $\pi_\sigma(w) = 2k + 1$.

Supposition: $\text{loc}(w) =: \ell < 2k + 1$

Let μ be an optimal marking sequence for w .

Additional supposition: μ not optimal for u

Then there exists a stage $i \in [|\text{alph}(w)|]$ such that $|\text{cond}(f(u_i))|_1 = k + 1$. If \mathbf{a} is unmarked in this stage, $|\text{cond}(f(w_i))|_1 = 2k + 2 > \ell$ which contradicts the first supposition. If \mathbf{a} is marked in this stage $|\text{cond}(f(w_i))|_1 = 2k + 1$ which contradicts the first supposition.

Thus, μ is optimal for u . By assumption there exists a stage $i \in [|\text{alph}(u)|]$ such that \mathbf{a} is marked, k blocks are marked, and $u[[u]]$ is unmarked. This implies since $\text{cond}(f(w_i))$ is a palindrome that at most $\frac{\ell-1}{2}$ blocks are marked in u . Thus, $k \leq \frac{\ell-1}{2} < \frac{2k+1-1}{2} = k$.

case 3. In the remaining case u is not rbpm, $\mathbf{a} \in \text{alph}(u)$, and there exists an optimal marking sequence for u such that in every stage \mathbf{a} is unmarked or less than k blocks are marked or $u[[u]]$ is marked. Let σ be such a marking sequence. Then $\pi_\sigma(w) = 2k$.

Supposition: $\text{loc}(w) =: \ell < 2k$

Let μ be an optimal marking sequence for w . Since u is not rbpm there exists a stage $i \in [|\text{alph}(u)|]$ such that $|\text{cond}(f(u_i))|_1 = k$ and $u[[u]]$ is unmarked. If \mathbf{a} were unmarked in stage i , $k = |\text{cond}(f(u_i))|_1 \leq \frac{\ell}{2} < k$ and if \mathbf{a} were marked in stage i , $k = |\text{cond}(f(u_i))|_1 \leq \frac{\ell-1}{2} < \frac{2k-1}{2} = k - \frac{1}{2}$. Thus $2k + 1 \leq \ell < 2k$ would hold. \square

Lemma 22. *Let $w = u^i$ be the i -times repetition for $u \in X^*$ and $i \in \mathbb{N}$. If u is strictly k -local then*

$$\text{loc}(w) = \begin{cases} ik - 1 + 1, & \text{if } u \text{ is bpm,} \\ ik, & \text{otherwise.} \end{cases}$$

Proof. Let σ be a marking sequence with $\pi_\sigma = \text{loc}(u) = k$. Since $\text{alph}(u) = \text{alph}(u^i)$ for all $i \in \mathbb{N}$, σ is also a marking sequence for w . If u is not bpm, there exists a stage during the marking in which k blocks are marked by σ and at least one of $u[1]$ or $u[|u|]$ is unmarked. Thus marking w according to the sequence σ leads to $\pi_\sigma(w) = ik$. If u is bpm, in any stage in which k blocks are marked, $u[1]$ and $u[|u|]$ are marked and thus in w , while being marked according to σ , the last marked block of an occurrence of u and the first marked block of the next occurrence of u coincide, as soon as the prefix of length $|u|$ of w contains k marked blocks. So, we get $\pi_\sigma(w) = ik - i + 1$.

For proving $\text{loc}(w) = ik$ or $\text{loc}(w) = ik - i + 1$ respectively, consider firstly $i = 2$. Assume first that w is bpm. Suppose $\text{loc}(w) = \ell < 2k - 1$. Let σ' be the marking sequence witnessing $\text{loc}(w) = \ell$. Since u is strictly k -local, there exists a stage in marking w by σ' in which u has k marked blocks. The second u has exactly as many marked blocks as the first one, so also k . In the best case, in w the last marked block of the first u and the first marked block of the second u are connected. Anyway, the number of marked blocks of w is, in that case, exactly $2k - 1$. A contradiction to the assumption $\text{loc}(w) = \ell < 2k - 1$. If u is not bpm, then, once again, there exists a stage in marking w by σ' in which u has k marked blocks. The second u has also exactly k marked block. But, in this case, in w the last marked block of the first u and the first marked block of the second u do not touch (as either the last letter of u or its first letter are not marked). So w has $2k$ marked blocks, a contradiction.

This reasoning can be trivially extended for $i > 2$. □

B A Many-One Reduction to Prove NP-Hardness of Loc

We use the following notations. Given a marking sequence σ , let σ^R be the marking sequence obtained by reversing σ (i.e. $\sigma^R(i) = \sigma(|X| - i + 1)$ for $1 \leq i \leq |X|$). We say that a marking sequence σ with $\pi_\sigma(\alpha) = \text{loc}(\alpha)$ is *near-optimal* (for α) if $\pi_\sigma(\alpha) \in \{\text{loc}(\alpha), \text{loc}(\alpha) + 1\}$.

The next lemma shows that, given two letters x_i, x_j of a word α , it is guaranteed that there exists a near-optimal marking sequence which marks x_i before x_j .

Lemma 23. *Let α be a word over the alphabet $X = \{x_1, x_2, \dots, x_n\}$. Let $\sigma : \{1, 2, \dots, |X|\} \rightarrow X$ be a marking sequence. Then $|\pi_\sigma(\alpha) - \pi_{\sigma^R}(\alpha)| \leq 1$.*

Proof. Let $1 \leq i \leq |X|$ and consider the marking of the first i letters in α according to σ^R . Note that these letters are exactly the last i letters to be marked according to σ . In particular, the number of marked blocks after stage i of marking α according to σ^R corresponds exactly to the number of unmarked blocks – or gaps – after stage $|X| - i$ of marking α according to σ . Since the number of unmarked blocks/gaps can be at most one higher, and at most one lower than the number of marked blocks, the lemma follows immediately. □

In this part of the appendix, we show, via a many-one reduction, that LOC is NP-hard. To this end, we devise a reduction from the well-known NP-complete CLIQUE problem, i.e., the problem to decide, for a given graph $\mathcal{G} = (V, E)$ and $\ell \in \mathbb{N}$, whether \mathcal{G} contains a clique (i.e., a complete subgraph) of size ℓ .

Let $\mathcal{G} = (V, E)$ be an undirected graph with $V = \{v_1, v_2, \dots, v_n\}$ and let $\ell \in \mathbb{N}$ with $\ell \leq n$. Note that the number of edges in a clique of size ℓ is exactly $\mu_\ell = \frac{\ell(\ell-1)}{2}$. We define the alphabet $X = \{x_1, x_2, \dots, x_n, z_1, z_2, z_3\}$ containing a unique letter for each vertex of the graph, along with three extra ‘control’ letters. Let $d(i)$ denote the degree of each vertex v_i , and let $\Delta = \max_{1 \leq i \leq n} \{d(i)\}$.

Next, we define the word $\alpha = \alpha_1 \alpha_2 \alpha_3$, where $\alpha_1 = (z_1 z_2 z_3 z_2)^{\gamma_1}$,

$$\alpha_2 = (z_1 z_2)^{\gamma_2(n-\ell)} (x_1 z_2)^{\gamma_2} (x_2 z_2)^{\gamma_2} \dots (x_n z_2)^{\gamma_2} (z_3 z_2)^{\ell \gamma_2} z_3,$$

$$\alpha_3 = \left(\prod_{\{v_i, v_j\} \in E \wedge i < j} (x_i x_j)^{\gamma_3} z_3 \right) \left(\prod_{1 \leq i \leq n} (x_i z_3)^{\gamma_3(\Delta - d(i))} \right),$$

and γ_1, γ_2 and γ_3 are chosen such that $\gamma_1 > |\alpha_2\alpha_3|$, $\gamma_2 > |\alpha_3| + 1$, and $\gamma_3 > 2$. Finally, let $\rho = \gamma_1 + n\gamma_2 + (\ell\Delta - 2\mu_\ell)\gamma_3 + \mu_\ell + 1$.

Lemma 24. *The word α is ρ -local if and only if \mathcal{G} contains a clique of size ℓ .*

Proof. We first consider some general observations on the k -locality of α . For clarity, and to avoid counting marked blocks more than once, we use the convention that a marked block which starts in α_1 and ends in α_2 (or α_3) belongs to α_2 (or α_3 , respectively), and *not* to α_1 .

Claim 1. α is $(\gamma_1 + n\gamma_2 + |\alpha_3|)$ -local.

Proof. (Claim 1) Consider the marking sequence $z_1, x_1, x_2, \dots, x_\ell, z_2, z_3, x_{\ell+1}, \dots, x_n$. After marking the first letter, z_1 , we have $\gamma_1 + \gamma_2(n - \ell)$ blocks. Marking the letters x_i , $1 \leq i \leq \ell$, introduces exactly $\ell\gamma_2$ additional blocks in α_2 (each single x_i accounting for γ_2 blocks), and altogether, they introduce fewer than $|\alpha_3|$ additional blocks in α_3 , resulting always in a total of less than $\gamma_1 + n\gamma_2 + |\alpha_3|$ blocks. Marking z_2 introduces no new blocks in α_1 (the last occurrence is adjacent to the first z_1 in α_2), and joins together $n\gamma_2$ blocks in α_2 while simultaneously introducing $n\gamma_2$ more, giving a net increase of one. Since z_2 does not occur in α_3 , no new blocks are introduced there. Thus we have at most $\gamma_1 + n\gamma_2 + |\alpha_3|$ blocks. Marking z_3 joins all the γ_1 blocks in α_1 , and α_1 is completely marked. Since no more than $|\alpha_2\alpha_3|$ blocks can exist elsewhere, and since $\gamma_1 > |\alpha_2\alpha_3|$, all further steps will have less than $\gamma_1 + 1$ marked blocks, so the maximum used is less than $\gamma_1 + n\gamma_2 + |\alpha_3| + 1$ as claimed. \square

Claim 2. In any optimal marking sequence, z_2 is marked between z_1 and z_3 . Consequently there exists a near-optimal marking sequence in which z_1 is marked before z_2 , which in turn is marked before z_3 .

Proof. (Claim 2) If z_2 were the first (resp. last) out of the three to be marked, then α_1 would contain $2\gamma_1 > \gamma_1 + n\gamma_2 + |\alpha_3|$ marked blocks and thus by Claim 1 the marking sequence is not optimal. The second statement follows from Lemma 1. \square

For the rest of the proof, consider a near-optimal marking sequence in which z_1 is marked before z_2 , and z_2 is marked before z_3 . Such a sequence exists, by Claim 2. Let ℓ' be the number of x_i s which are marked before z_2 . If $\ell' < \ell$, then exactly after z_2 is marked, we have γ_1 marked blocks in α_1 . The number of marked blocks in the suffix $(z_3z_2)^{\ell\gamma_2}z_3$ of α_2 is $\ell\gamma_2$. To count the marked blocks in the rest of α_2 (i.e. the prefix $(z_1z_2)^{\gamma_2(n-\ell)}(x_1z_2)^{\gamma_2}(x_2z_2)^{\gamma_2} \dots (x_nz_2)^{\gamma_2}$), note that since both ends of this factor are marked, the number of marked blocks is exactly one more than the number of gaps. Moreover, since z_1 and z_2 are marked, the only gaps come from occurrences of the unmarked x_i s. Since no two occurrences of these are adjacent, this means that each occurrence is a unique gap so there are $(n - \ell')\gamma_2$ gaps in total. Consequently, α_2 contains exactly $(n - \ell' + \ell)\gamma_2 + 1$ marked blocks. Since $\gamma_2(n - \ell' + \ell) \geq \gamma_2(n + 1)$, and $\gamma_2 > |\alpha_3| + 1$, this means we have more than $\gamma_1 + n\gamma_2 + |\alpha_3| + 1$ blocks in total. By Claim 1, this contradicts our assumption that the sequence is near optimal. Similarly, if $\ell' > \ell$, then exactly before z_2 is marked, we have γ_1 marked blocks in α_1 and $\gamma_2(n - \ell + \ell') \geq \gamma_2(n + 1)$ marked blocks in α_2 . Again this implies that we have more than $\gamma_1 + n\gamma_2 + |\alpha_3| + 1$ marked blocks altogether, contradicting the assumption that our sequence is near-optimal. Consequently, $\ell' = \ell$, and there exist i_1, i_2, \dots, i_ℓ such that the set of letters marked before z_2 is $\{x_{i_1}, x_{i_2}, \dots, x_{i_\ell}, z_1\}$.

Now, we observe that after z_2 is marked, the number of marked blocks is never increased. To see why, suppose $z_1, z_2, x_{i_1}, x_{i_2}, \dots, x_{i_\ell}$ are marked (note that we do not exclude the case that more letters may also be marked). Suppose we mark z_3 . Then γ_1 marked blocks will be joined together in α_1 , and hence decrease the number of marked blocks in α_1 by $\gamma_1 - 1$ (and α_1 is completely marked). Since $\gamma_1 > |\alpha_2\alpha_3|$, the total number of blocks cannot increase overall. Similarly, suppose we mark some x_j , $1 \leq j \leq n$. Then γ_2 marked blocks are joined together in α_2 , thus reducing the number of marked blocks by $\gamma_2 - 1$. The number of blocks in α_1 remains the same, and since $\gamma_2 > |\alpha_3|$, the total number of marked blocks cannot increase overall.

It is reasonably straightforward to observe that until z_2 is marked, the total number of marked blocks is never decreased (in order to be fully precise, one can make an argument symmetric to the above). Thus, the maximum number of marked blocks in our sequence is obtained (not necessarily for the first time) when z_2 is marked. In other words, if exactly $z_1, z_2, x_{i_1}, x_{i_2}, \dots, x_{i_\ell}$ are marked, we have the maximal number of blocks. Clearly, this implies that there are γ_1 blocks in α_1 and $n\gamma_2 + 1$ blocks in α_2 .

We now consider the number of marked blocks in α_3 , which is given by $\gamma_3(\Delta\ell - 2t) + t$, where $t = |\{(j, j') \mid 1 \leq j < j' \leq \ell \wedge \{v_{i_j}, v_{i_{j'}}\} \in E\}|$. To see this, first suppose there are gaps (or a new unmarked letter #) between all adjacent letters. This hypothetical situation would give a total of $\Delta\gamma_3\ell$ blocks. Then consider how many blocks are lost or joined by removing the gaps (or #s). In particular, precisely $2\gamma_3$ blocks are joined together for each pair $x_{i_j}, x_{i_{j'}}$ such that $\{v_{i_j}, v_{i_{j'}}\} \in E$. No further blocks are joined together - so for each such pair we must subtract $2\gamma_3 - 1$ from the total.

Note that t can be at most μ_ℓ and is exactly μ_ℓ if and only if the vertices $v_{i_1}, v_{i_2}, \dots, v_{i_\ell}$ form a clique. Consequently, if G contains a size- ℓ clique, a (near-optimal) marking sequence can be chosen such that the maximum number of blocks used is $\gamma_1 + n\gamma_2 + \gamma_3(\Delta\ell - 2\mu_\ell) + \mu_\ell + 1 = \rho$. Hence, in this case, α is ρ -local. On the other hand, if G does not contain a size- ℓ clique, then regardless of the choice of $x_{i_1}, x_{i_2}, \dots, x_{i_\ell}$, we have $t \leq \mu_\ell - 1$, and any near optimal marking sequence requires at least

$$\gamma_1 + n\gamma_2 + \gamma_3(\Delta\ell - 2\mu_\ell + 2) + \mu_\ell = \gamma_1 + n\gamma_2 + (\Delta\ell - 2\mu_\ell)\gamma_3 + 2\gamma_3 + \mu_\ell > \rho$$

marked blocks, meaning α is not ρ -local. Thus α is ρ -local if and only if G contains a size- ℓ clique. Since α and ρ can be constructed in polynomial time, the theorem follows. \square

Since LOC is obviously in NP, Lemma 24 leads to an alternative proof that Theorem 6 holds.

C Investigation of Simple Greedy Strategies for Locality Number

We shall formulate the greedy strategies in a bit more detail:

FewOcc	Among all unmarked symbols, choose one with a smallest number of occurrences.
ManyOcc	Among all unmarked symbols, choose one with a largest number of occurrences.
FewBlocks	Among all unmarked symbols, choose one that, after marking it, results in the smallest total number of marked blocks.
LeftRight	Among all unmarked symbols, choose the one with the leftmost occurrence.
BlockExt	Among all unmarked symbols, choose one that has at least one occurrence that is adjacent to a marked block.

These strategies are – except for **LeftRight** – nondeterministic, since there are in general several valid choices of the next symbol to mark. However, showing poor performances independent of the nondeterministic choices are stronger negative results. We make the convention that all strategies – except, of course, **LeftRight** – can choose any symbol as the initially marked symbol, which is justified by the fact that, in terms of running-time, we could afford to try out every possible choice of the first symbol. In the following, for every greedy strategy S and for every word α , let $\text{GREEDY}_S(\alpha)$ be the optimal marking number over all marking sequences that can be obtained by strategy S , let $\psi_S(\alpha) = \frac{\text{GREEDY}_S(\alpha)}{\text{loc}(\alpha)}$ and $\psi_S = \max_\alpha \{\psi_S(\alpha)\}$.

Proposition 25. *Let $\ell \geq 2$. Then $\psi_{\text{BlockExt}}(x_1y x_2y x_3y \dots x_\ell y) \geq 2 - \frac{2}{\ell}$.*

Proof. For the sake of convenience, let $\ell = 2k$ for some $k \geq 1$. Assume that α is marked by greedy strategy **BlockExt**. If y is marked first, we have $2k$ marked blocks and if some x_i , $1 \leq i \leq 2k$, is marked first, then y is marked next, which leads to $2k - 1$ marked blocks. From now on, marking the rest of the symbols decreases the number of marked blocks; thus, $\text{GREEDY}_{\text{BlockExt}}(\alpha) = 2k - 1$. On the other hand, if we do not stick to strategy **BlockExt**, then we can first mark the k symbols x_2, x_3, \dots, x_{k+1} , which leads to k marked blocks. Then marking y joins all the previously marked blocks into one marked block and turns $k - 1$ occurrences of y into new individual marked blocks (i. e., the $k - 2$ occurrences of y between the symbols $x_{k+2}, x_{k+3}, \dots, x_{2k}$ and the single occurrence of y after x_{2k}). Thus, there are k marked blocks. Since from now on marking the rest of the symbols only decreases the number of marked blocks, we conclude that $\text{loc}(\alpha) \leq k$. Consequently, $\psi_{\text{BlockExt}}(\alpha) \geq \frac{\ell-1}{2} = 2 - \frac{2}{\ell}$. \square

For every $S \in \{\text{FewOcc}, \text{ManyOcc}, \text{FewBlocks}, \text{LeftRight}\}$, by $\text{BlockExt} - S$, we denote the strategy BlockExt with the addition that strategy S is applied in order to decide between symbols that are valid candidates with respect to BlockExt , i. e., symbols that extend at least one marked block (for $\text{BlockExt} - \text{LeftRight}$ this means that among all valid candidates we choose the one that occurs as the leftmost).

Let $\ell \geq 2$ and let

$$\begin{aligned} \alpha &= (x_1 x_2 \dots x_\ell)^2 x_1 \beta_1 x_2 \beta_2 x_3 \beta_3 \dots \beta_{\ell-1} x_\ell, \\ &\text{where, for every } i, 1 \leq i \leq \ell - 1, \beta_i = (y_{2i-1} y_{2i})^4, 1 \leq i \leq \ell - 1, \\ \gamma &= x_1 x_2 \dots x_\ell x_1 y_1 x_2 y_2 x_3 y_3 \dots y_{\ell-1} x_\ell. \end{aligned}$$

Proposition 26. *For every $S \in \{\text{FewOcc}, \text{FewBlocks}\}$, $\psi_{\text{BlockExt} - S}(\alpha) \geq \frac{\ell-1}{6}$, $\psi_S(\alpha) \geq \frac{\ell-1}{6}$, $\psi_{\text{BlockExt} - \text{ManyOcc}}(\gamma) \geq \frac{\ell-1}{6}$ and $\psi_{\text{ManyOcc}}(\gamma) \geq \frac{\ell-1}{2}$.*

Proof. We first consider α and observe that $(x_1, y_1, y_2, x_2, y_3, y_4, x_3, y_5, y_6, \dots)$ is an optimal marking sequence which shows that $\text{loc}(\alpha) = 6$. Next, we consider how the strategies $\text{BlockExt} - S$, $S \in \{\text{FewOcc}, \text{FewBlocks}\}$, can mark α . If the first marked symbol is some x_i , then both $\text{BlockExt} - \text{FewOcc}$ and $\text{BlockExt} - \text{FewBlocks}$ would next mark all remaining x_j , $j \neq i$, in such an order that there are always at most 2 marked blocks in the prefix $(x_1 x_2 \dots x_\ell)^2$. This leads to at least ℓ marked blocks. If, on the other hand, some y_{2i-1} or y_{2i} is marked first, then $\text{BlockExt} - \text{FewOcc}$ marks x_i or x_{i+1} next (depending on whether y_{2i-1} or y_{2i} is marked first) and then all remaining x_j as before, while $\text{BlockExt} - \text{FewBlocks}$ would mark the remaining symbol of y_{2i-1} or y_{2i} and then all x_j . This results in at least $\ell - 1$ marked blocks. Thus, $\psi_{\text{BlockExt} - S}(\alpha) \geq \frac{\ell-1}{6}$, $S \in \{\text{FewOcc}, \text{FewBlocks}\}$. Moreover, strategies FewOcc and FewBlocks will behave on α just like $\text{BlockExt} - \text{FewOcc}$ and $\text{BlockExt} - \text{FewBlocks}$, respectively, with the only difference that they have more freedom with respect to the order of how the x_i are marked. In particular, $\psi_S(\alpha) \geq \frac{\ell-1}{6}$, $S \in \{\text{FewOcc}, \text{FewBlocks}\}$.

Next, we consider the word γ and observe that $(x_1, y_1, x_2, y_2, x_3, y_3, \dots)$ is an optimal marking sequence which shows that $\text{loc}(\gamma) = 2$. If $\text{BlockExt} - \text{ManyOcc}$ marks some x_i first, then it will mark all remaining x_j next, which results in ℓ marked blocks. If, on the other hand, the first symbol is some y_i , then either x_i or x_{i+1} is marked next and after that all remaining x_j , leading to $\ell - 1$ marked blocks. The strategy ManyOcc behaves in the same way, with the only difference that there are more possibilities in which order the symbols x_i are marked. Thus, $\psi_S(\gamma) \geq \frac{\ell-1}{2}$. \square

Let $\ell \geq 2$ be an even number and let

$$\delta = x_1 x_2 \dots x_\ell x_1 x_\ell x_2 x_{\ell-1} x_3 x_{\ell-2} \dots x_{\frac{\ell}{2}} x_{\frac{\ell}{2}+1}.$$

Proposition 27. *Then $\psi_{\text{BlockExt} - \text{LeftRight}}(\delta) \geq \frac{\ell}{4}$ and $\psi_{\text{LeftRight}}(\delta) \geq \frac{\ell}{4}$.*

Proof. We first observe that $\text{loc}(\delta) = 2$, which is witnessed by the marking sequence

$$(x_1, x_\ell, x_2, x_{\ell-1}, x_3, x_{\ell-2}, \dots, x_{\frac{\ell}{2}}, x_{\frac{\ell}{2}+1})$$

(note that this marking sequence maintains a marked prefix and one additional marked internal factor starting with $x_\ell x_1 x_\ell$, which is alternately extended to both sides).

Assume that x_i is the first symbol marked by $\text{BlockExt} - \text{LeftRight}$. If $i \leq \frac{\ell}{2}$, then we mark next x_{i-1} , then x_{i-2} and so on, until all x_1, x_2, \dots, x_i are marked. Then the symbols $x_{i+1}, x_{i+2}, \dots, x_{\frac{\ell}{2}}$ are marked, which leads to $\frac{\ell}{2} + 1$ marked blocks. If, on the other hand, $i = \frac{\ell}{2} + j$ with $j \geq 1$, then the next marked symbol will be $x_{\frac{\ell}{2} - (j-1)}$. Then, as before, we will mark $x_{\frac{\ell}{2} - (j-1) - 1}, x_{\frac{\ell}{2} - (j-1) - 2}, \dots$ until all $x_1, x_2, \dots, x_{\frac{\ell}{2} - (j-1)}$ are marked, and then $x_{\frac{\ell}{2} - (j-1) + 1}, \dots, x_{\frac{\ell}{2}}$ are marked. This results in $\frac{\ell}{2}$ marked blocks. Thus, $\psi_{\text{BlockExt} - \text{LeftRight}}(\delta) \geq \frac{\ell}{4}$. If the strategy LeftRight marks some symbol x_i with $i \leq \frac{\ell}{2}$ first, then it marks next all the symbol $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_{\frac{\ell}{2}}$, which results in $\frac{\ell}{2} + 1$ marked blocks. If, on the other hand, $i > \frac{\ell}{2}$, then symbols $x_1, \dots, x_{\frac{\ell}{2}}$ are marked next, which leads to at least $\frac{\ell}{2}$ marked blocks. Thus $\psi_{\text{LeftRight}}(\delta) \geq \frac{\ell}{4}$. \square

We define the following extensions of BlockExt:

BlockExt-1 Among all extending symbols, choose one that has the most extending occ.

BlockExt-2 Among all extending symbols, choose one for which $\frac{\# \text{extending occ.}}{\# \text{occ.}}$ is maximal.

Proposition 28. *For every $S \in \{\text{BlockExt-1}, \text{BlockExt-2}\}$, $\psi_S(\alpha) \geq \frac{\ell-1}{6}$.*

Proof. If we first mark a symbol x_i , then, among all symbols extending a marked block, i. e., symbols x_{i-1} , x_{i+1} , y_{2i-1} and y_{2i+1} , the symbols x_{i-1} and x_{i+1} each have 3 occurrences in total, two of which are extending a marked block, whereas the symbols y_{2i-1} and y_{2i+1} each have 4 occurrences, only one of which is extending. Consequently, both BlockExt-1 and BlockExt-2 chose either x_{i-1} and x_{i+1} next. This situation does not change until all x_i are marked, which leads to ℓ marked blocks. If, on the other hand, some y_{2i-1} or y_{2i} is marked first, then we mark next the remaining symbol y_{2i-1} or y_{2i} such that β_i is completely marked. Next, x_i and x_{i+1} are marked, in some order, which brings us back to the situation described above which leads to the marking of all remaining x_j , leading to $\ell - 1$ marked blocks. Consequently, for every $S \in \{\text{BlockExt-1}, \text{BlockExt-2}\}$, $\psi_S(\alpha) \geq \frac{\ell-1}{6}$ (note that $\text{loc}(\alpha) \leq 6$ is discussed in the proof of Proposition 26). \square