On the Readability of Overlap Digraphs

Rayan Chikhi\textsuperscript{a,b}, Paul Medvedev\textsuperscript{a,d,e,1}, Martin Milani\textsuperscript{c}, Sofya Raskhodnikova\textsuperscript{a}

\textsuperscript{a}Department of Computer Science, The Pennsylvania State University, USA
\textsuperscript{b}CNRS, UMR 9189, France
\textsuperscript{c}UP IAM and UP FAMNIT, University of Primorska, Slovenia
\textsuperscript{d}Department of Biochemistry and Molecular Biology, The Pennsylvania State University, USA
\textsuperscript{e}Genome Sciences Institute of the Huck, The Pennsylvania State University, USA

Abstract

We introduce the graph parameter \textit{readability} and study it as a function of the number of vertices in a graph. Given a digraph $D$, an injective overlap labeling assigns a unique string to each vertex such that there is an arc from $x$ to $y$ if and only if $x$ properly overlaps $y$. The readability of $D$ is the minimum string length for which an injective overlap labeling exists. In applications that utilize overlap digraphs (e.g., in bioinformatics), readability reflects the length of the strings from which the overlap digraph is constructed. We study the asymptotic behaviour of readability by casting it in purely graph theoretic terms (without any reference to strings). We prove upper and lower bounds on readability for certain graph families and general graphs.

\textit{Keywords:} graph parameter, stringology, bioinformatics, readability, overlap graph

1. Introduction

In this paper, we introduce and study a graph parameter called readability, motivated by applications of overlap graphs in bioinformatics. A string $x$ \textit{overlaps} a string $y$ if there is a suffix of $x$ that is equal to a prefix...
of \( y \). They overlap properly if, in addition, the suffix and prefix are both proper. The overlap digraph of a set of strings \( S \) is a digraph where each string is a vertex and there is an arc from \( x \) to \( y \) (possibly with \( x = y \)) if and only if \( x \) properly overlaps \( y \). Walks in the overlap digraph of \( S \) represent strings that can be spelled by stitching strings of \( S \) together, using the overlaps between them. Overlap digraphs have various applications, e.g., they are used by approximation algorithms for the Shortest Superstring Problem [Swe00]. Their most impactful application, however, has been in bioinformatics. Their variants, such as de Bruijn graphs [IW95] and string graphs [Mye05], have formed the basis of nearly all genome assemblers used today (see [MKS10, NP13] for a survey), successful despite results showing that assembly is a hard problem in theory [BBT13, NP09, MGMB07]. In this context, the strings of \( S \) represent known fragments of the genome (called reads), and the genome is represented by walks in the overlap digraph of \( S \). However, do the overlap digraphs generated in this way capture all possible digraphs, or do they have any properties or structure that can be exploited?

Braga and Meidanis [BM02] showed that overlap digraphs capture all possible digraphs, i.e., for every digraph \( D \), there exists a set of strings \( S \) such that their overlap digraph is \( D \). Their proof takes an arbitrary digraph and shows how to construct an injective overlap labeling, that is, a function assigning a unique string to each vertex, such that \((x, y)\) is an arc if and only if the string assigned to \( x \) properly overlaps the string assigned to \( y \). However, the length of strings produced by their method can be exponential in the number of vertices. In the bioinformatics context, this is unrealistic, as the read size is typically much smaller than the number of reads.

To investigate the relationship between the string length and the number of vertices, we introduce a graph parameter called readability. The readability of a digraph \( D \), denoted \( r(D) \), is the smallest nonnegative integer \( r \) such that there exists an injective overlap labeling of \( D \) with strings of length \( r \). The result by [BM02] shows that readability is well defined and is at most \( 2^{\Delta + 1} - 1 \), where \( \Delta \) is the maximum of the in- and out-degrees of vertices in \( D \). However, nothing else is known about the parameter, though there are papers that look at related notions [BFK+02, BFKK02, BHKdW99, GP14, LZ07, LZ10, PSW03, TU88].

In this paper, we study the asymptotic behaviour of readability as a function of the number of vertices in a graph. We define readability for undirected bipartite graphs and show that the two definitions of readability are asymptotically equivalent. We capture readability using purely graph theoretic parameters (i.e., without any reference to strings). For trees, we give a parameter that characterizes readability exactly. For the larger family
of bipartite $C_4$-free graphs, we give a parameter that approximates readability to within a factor of 2. Finally, for general bipartite graphs, we give a parameter that is bounded on the same sets of graphs as readability.

We apply our purely graph theoretic interpretation to prove readability upper and lower bounds on several graph families. We show, using a counting argument, that almost all digraphs and bipartite graphs have readability of at least $\Omega(n/\log n)$. Next, we construct a graph family inspired by Hadamard codes and prove that it has readability $\Omega(n)$. Finally, we show that the readability of trees is bounded from above by their radius, and there exist trees of arbitrary readability that achieve this bound.

2. Preliminaries

2.1. General definitions and notation.

We use $\epsilon$ to denote the empty string. Let $x$ be a string. We denote the length of $x$ by $|x|$. We use $x[i]$ to refer to the $i^{th}$ character of $x$, and denote by $x[i..j]$ the substring of $x$ from the $i^{th}$ to the $j^{th}$ character, inclusive. We let $\text{pre}_i(x)$ denote the prefix $x[1..i]$ of $x$, and we let $\text{suf}_i(x)$ denote the suffix $x[|x| - i + 1..|x|]$. Let $y$ be another string. We denote by $x \cdot y$ the concatenation of $x$ and $y$. We say that $x$ overlaps $y$ if there exists an $i$ with $1 \leq i \leq \min\{|x|, |y|\}$ such that $\text{suf}_i(x) = \text{pre}_i(y)$. In this case, we say that $x$ overlaps $y$ by $i$. If $i < \min\{|x|, |y|\}$, then we call the overlap proper. Define $\text{ov}(x, y)$ as the minimum $i$ such that $x$ overlaps $y$ by $i$, or 0 if $x$ does not overlap $y$. For a positive integer $n$, we denote by $[n]$ the set $\{1, \ldots, n\}$.

We refer to finite simple undirected graphs simply as graphs and to finite directed graphs without parallel arcs in the same direction as digraphs. For a vertex $v$ in a graph, we denote the set of neighbors of $v$ by $N(v)$. A $P_4$ denotes the path on 4 vertices and 3 edges. A biclique is a complete bipartite graph. Note that the one-vertex graph is a biclique (with one of the parts of its bipartition being empty). Two vertices $u, v$ in a graph are called twins if they have the same neighbors, i.e., if $N(u) = N(v)$. If, in addition, $N(u) = N(v) \neq \emptyset$, vertices $u, v$ are called non-isolated twins. A matching is a graph of maximum degree at most 1, though we will sometimes slightly abuse the terminology and not distinguish between matchings and their edge sets. A cycle (respectively, path) on $i$ vertices is denoted by $C_i$ (respectively, $P_i$). For graph terms not defined here, see, e.g., [BM08].

Define $B_{n \times n}$ as the set of balanced bipartite graphs with nodes $[n]$ in each part, and define $D_n$ as the set of all digraphs with nodes $[n]$. 
2.2. Readability of digraphs.

A labeling \( \ell \) of a graph or digraph is a function assigning a string to each vertex such that all strings have the same length, denoted by \( \text{len}(\ell) \). We define \( \text{ov}_\ell(u, v) = \text{ov}(\ell(u), \ell(v)) \). An overlap labeling of a digraph \( D = (V, A) \) is a labeling \( \ell \) such that \((u, v) \in A \text{ if and only if } 0 < \text{ov}_\ell(u, v) < \text{len}(\ell) \). An overlap labeling is said to be injective if it does not generate duplicate strings. Recall that the readability of a digraph \( D \), denoted \( r(D) \), is the smallest nonnegative integer \( r \) such that there exists an injective overlap labeling of \( D \) of length \( r \). We note that in our definition of readability we do not place any restrictions on the alphabet size. Braga and Meidanis [BM02] gave a reduction from an overlap labeling of length \( \ell \) over an arbitrary alphabet \( \Sigma \) to an overlap labeling of length \( \ell \log |\Sigma| \) over the binary alphabet.

2.3. Readability of bipartite graphs.

We also define a modified notion of readability that applies to balanced bipartite graphs as opposed to digraphs. We found that readability on balanced bipartite graphs is simpler to study but is asymptotically equivalent to readability on digraphs. Let \( G = (V, E) \) be a bipartite graph with a given bipartition of its vertex set \( V(G) = V_s \cup V_p \). (We also use the notation \( G = (V_s, V_p, E) \).) We say that \( G \) is balanced if \(|V_s| = |V_p| \). An overlap labeling of \( G \) is a labeling \( \ell \) of \( G \) such that for all \( u \in V_s \) and \( v \in V_p \), \((u, v) \in E \) if and only if \( \text{ov}_\ell(u, v) > 0 \). In other words, overlaps are exclusively between the suffix of a string assigned to a vertex in \( V_s \) and the prefix of a string assigned to a vertex in \( V_p \). The readability of \( G \) is the smallest nonnegative integer \( r \) such that there exists an overlap labeling of \( G \) of length \( r \). Note that we do not require injectivity of the labeling, nor do we require the overlaps to be proper. As before, we use \( r(G) \) to denote the readability of \( G \).

We note that, in our definition of readability, we do not place any restrictions on the alphabet size. Braga and Meidanis [BM02] gave a reduction from an overlap labeling of length \( \ell \) over an arbitrary alphabet \( \Sigma \) to an overlap labeling of length \( \ell \log |\Sigma| \) over the binary alphabet.

For a labeling \( \ell \), we define \( \text{inner}_i(\ell(v)) = \text{suf}_i(\ell(v)) \) if \( v \in V_s \) and \( \text{inner}_i(\ell(v)) = \text{pre}_i(\ell(v)) \) if \( v \in V_p \). Similarly, we define \( \text{outer}_i(\ell(v)) = \text{pre}_i(\ell(v)) \) if \( v \in V_s \) and \( \text{outer}_i(\ell(v)) = \text{suf}_i(\ell(v)) \) if \( v \in V_p \).

3. Relationship of readability of bipartite graphs and digraphs

In this section, we show that the readabilities of digraphs and of bipartite graphs are asymptotically equivalent. More precisely, we will prove the
following theorem.

**Theorem 3.1.** There exists a bijection \( \psi : \mathcal{B}_{n \times n} \to \mathcal{D}_n \) with the property that for any \( G \in \mathcal{B}_{n \times n} \) and \( D \in \mathcal{D}_n \), such that \( D = \psi(G) \), we have that \( r(G) < r(D) \leq 2 \cdot r(G) + 1 \).

As a result, we can study readability of balanced bipartite graphs, without asymptotically affecting our bounds. For example, we show in Section 5.2 (in Theorem 5.2) that there exists a family of balanced bipartite graphs with readability \( \Omega(n) \), which leads to the existence of digraphs with readability \( \Omega(n) \). In the rest of the section, we will prove Theorem 3.1.

To disambiguate the two partitions of a bipartite graph, we label the vertices of \( G = (V_s, V_p, E) \in \mathcal{B}_{n \times n} \) using notation \( V_s = \{i_s \mid i \in [n]\} \) and \( V_p = \{i_p \mid i \in [n]\} \). For the proof, we define the following transformation. Let \( D = ([n], A) \in \mathcal{D}_n \). Define \( \phi(D) = (V_s, V_p, E) \) as the bipartite graph with \( V_s = \{i_s \mid i \in [n]\} \), \( V_p = \{i_p \mid i \in [n]\} \). and \( E = \{(i_s, j_p) \mid (i, j) \in A\} \). This transformation was proposed in [BM02]. Similarly, we define the transformation \( \psi \), as follows. Given a bipartite graph \( G = (V_s, V_p, E) \in \mathcal{B}_{n \times n} \), we define \( \psi(G) = ([n], A) \) where \( A = \{(i, j) \mid (i_s, j_p) \in E\} \). It is easy to see that \( \psi \) is a bijection from \( \mathcal{B}_{n \times n} \) to \( \mathcal{D}_n \), as required, and \( \phi \) is its inverse.

The following two lemmas prove the readability bounds stated in the theorem.

**Lemma 3.1.** Let \( D = (V, A) \in \mathcal{D}_n \) be a digraph with \( A \neq \emptyset \). Then \( r(\phi(D)) < r(D) \).

_Proof._ Let \( \ell \) be an injective overlap labeling of \( D \). Since \( A \neq \emptyset \), we have \( \text{len}(\ell) \geq 1 \). Define a labeling \( \ell_{\phi} \) of \( \phi(D) \) as follows. For \( w \in V \), let \( \ell_{\phi}(w_s) = \ell(w)[2..|\ell(w)|] \) and let \( \ell_{\phi}(w_p) = \ell(w)[1..|\ell(w)| - 1] \). (If \( |\ell(w)| = 1 \), then each of \( \ell_{\phi}(w_s) \) and \( \ell_{\phi}(w_p) \) is the empty string.) It is clear that \( \ell_{\phi} \) is a labeling of \( \phi(D) \) of length \( \text{len}(\ell) - 1 \). We claim that \( \ell_{\phi} \) is an overlap labeling of \( \phi(D) \). Suppose that \( (u_s, v_p) \in E(\phi(D)) \). Then \( (u, v) \in A \), which implies \( \text{ov}_\ell(u, v) > 0 \). Also, \( \text{ov}_\ell(u, v) < \text{len}(\ell) \). Consequently, the shortest overlap between \( \ell(u) \) and \( \ell(v) \) yields an overlap between \( \ell_{\phi}(u_s) \) and \( \ell_{\phi}(v_p) \), implying \( \text{ov}_{\ell_{\phi}}(u_s, v_p) > 0 \). Conversely, the condition \( \text{ov}_{\ell_{\phi}}(u_s, v_p) > 0 \) implies \( 0 < \text{ov}_\ell(u, v) < \text{len}(\ell) \). Therefore, \( (u, v) \in A \) and, by the definition of \( \phi(D) \), also \( (u_s, v_p) \in E(\phi(D)) \). This shows that \( r(\phi(D)) \leq r(D) - 1 \). \qed

**Lemma 3.2.** Let \( G = (V_s, V_p, E) \in \mathcal{B}_{n \times n} \). Then \( r(\psi(G)) \leq 2 \cdot r(G) + 1 \).

_Proof._ Let \( \ell_G \) be an overlap labeling of \( G \) and let \( D = (V, A) = \psi(G) \), with \( V = [n] \). For \( w \in V \), define \( \ell(w) = \ell_G(w_p) \cdot w \cdot \ell_G(w_s) \). Here, \( w \) is treated as
a character in the alphabet $[n]$. We assume without loss of generality that these characters are distinct from the alphabet over which $\ell_G$ is defined. It is clear that $\ell$ is a labeling of $D$ of length $2 \cdot \text{len}(\ell_G) + 1$. We claim that $\ell$ is an injective overlap labeling of $D$. For every vertex $w \in V$, its label contains a distinct middle character corresponding to $w$, which implies injectivity. Now, suppose that $(u, v) \in A$. Then $(u_s, v_p) \in E$, which implies $\text{ov}_{\ell_G}(u_s, v_p) > 0$. By construction of $\ell$, it follows that $0 < \text{ov}_{\ell}(u, v) \leq \text{len}(\ell_G) < \text{len}(\ell)$. Conversely, suppose that $\text{ov}_{\ell}(u, v) > 0$. By construction of $\ell$, it follows that $\text{ov}_{\ell}(u, v) \leq \text{len}(\ell_G)$. Therefore, $\text{ov}_{\ell_G}(u_s, v_p) = \text{ov}_{\ell}(u, v) > 0$, which implies $(u_s, v_p) \in E$ and consequently $(u, v) \in A$. This shows that $r(\psi(G)) \leq 2 \cdot r(G) + 1$. 

Given $G \in B_{n \times n}$, we can apply the two lemmas to derive the inequality stated in Theorem 3.1:

$$r(G) = r(\phi(\psi(G))) < r(\psi(G)) \leq 2 \cdot r(G) + 1.$$ 

4. Graph theoretic characterizations

In this section, we relate readability of balanced bipartite graphs to several purely graph theoretic parameters, without reference to strings.

4.1. Trees and $C_4$-free graphs

For trees, we give an exact characterization of readability, while for $C_4$-free graphs, we give a parameter that is a 2-approximation to readability. A decomposition of size $k$ of a bipartite graph $G = (V_s, V_p, E)$ is a function on the edges of the form $w : E \rightarrow [k]$. Note that a labeling $\ell$ of $G$ implies a decomposition of $G$, defined by $w(e) = \text{ov}_{\ell}(e)$ for all $e \in E$. We call this the $\ell$-decomposition. We say that a labeling $\ell$ of $G$ achieves $w$ if it is an overlap labeling and $w$ is the $\ell$-decomposition. Note that we can express readability as

$$r(G) = \min\{k \mid w \text{ is a decomposition of size } k, \exists \text{ a labeling } \ell \text{ that achieves } w\}.$$ 

Our goal is to characterize in graph theoretic terms the properties of $w$ which are satisfied if and only if $w$ is the $\ell$-decomposition, for some $\ell$. While this is still open in general, we can achieve this for trees using a condition which we call the $P_4$-rule. We say that $w$ satisfies the $P_4$-rule if for every induced four-vertex path $P = (e_1, e_2, e_3)$ in $G$, the following condition holds: if $w(e_2) = \max\{w(e_1), w(e_2), w(e_3)\}$, then $w(e_2) \geq w(e_1) + w(e_3)$. The following theorem states our characterization of readability for trees.
Theorem 4.1. Let $T$ be a tree. Then $r(T) = \min \{ k \mid w \text{ is a decomposition of size } k \text{ that satisfies the } P_4\text{-rule} \}$.

Note that for cycles, the equality does not hold. For example, consider the decomposition $w$ of $C_6$ given by the weights $2, 4, 2, 2, 3, 1$. This decomposition satisfies the $P_4$ rule but it can be shown using case analysis that there does not exist a labeling $\ell$ achieving $w$.

However, we can give a characterization of readability for $C_4$-free graphs in terms of a parameter that is asymptotically equivalent to readability, using a condition which we call the strict $P_4$-rule. The strict $P_4$-rule is identical to the $P_4$-rule accept that the inequality becomes strict. That is, $w$ satisfies the strict $P_4$-rule if for every induced four-vertex path $P = (e_1, e_2, e_3)$, if $w(e_2) = \max \{w(e_1), w(e_2), w(e_3)\}$, then $w(e_2) > w(e_1) + w(e_3)$.

Note that a decomposition that satisfies the strict $P_4$-rule automatically satisfies the $P_4$-rule, but not vice-versa. The following theorem gives a 2-approximation of readability for $C_4$-free graphs.

Theorem 4.2. Let $G$ be a $C_4$-free bipartite graph. Let $t = \min \{ k \mid w \text{ is a decomposition of size } k \text{ that satisfies the strict } P_4\text{-rule} \}$. Then $t/2 < r(G) \leq t$.

We note that this characterization cannot be extended to graphs with a $C_4$. The example in Figure 1 shows a graph with a decomposition which satisfies the strict $P_4$-rule but it can be shown using case analysis that there does not exists a labeling $\ell$ achieving this decomposition.

In the remainder of this section, we prove Theorems 4.1 and 4.2. We first show that an $\ell$-decomposition satisfies the $P_4$-rule.
Lemma 4.1. Let $\ell$ be an overlap labeling of a bipartite graph $G$. Then the $\ell$-decomposition satisfies the $P_4$-rule.

Proof. Let $G = (V_s, V_p, E)$. Denote by $\ell$ be the $\ell$-decomposition. Suppose for the sake of contradiction that $w$ violates the $P_4$-rule. Then, there exists an induced four-vertex path $P = (u_1, u_2, u_3, u_4)$ in $G$ with $u_1 \in V_p$ (and consequently $u_2, u_4 \in V_s$ and $u_3 \in V_p$) such that
\[
\max\{w(v_1, v_2), w(v_2, v_3), w(v_3, v_4)\} = w(v_2, v_3) < w(v_1, v_2) + w(v_3, v_4).
\]
Then, $b = \max\{a, b, c\}$ and $b < a + c$, where $a = ov_1(u_2, u_1)$, $b = ov_1(u_2, u_3)$, and $c = ov_1(u_4, u_3)$. We will show that there exists an overlap from $\ell(u_1)$ to $\ell(u_4)$ of length $a + c - b$, which will prove the lemma, by contradicting the fact that $\ell$ is an overlap labeling and $(u_4, u_1) \notin E$ (as $P$ is an induced $P_4$).

Let $\ell$ be the length of $\ell$. Writing the overlaps in terms of substrings, we obtain that $\text{sup}_a(\ell(u_2)) = \text{pre}_a(\ell(u_1))$, $\text{sup}_b(\ell(u_2)) = \text{pre}_b(\ell(u_3))$, and $\text{sup}_c(\ell(u_4)) = \text{pre}_c(\ell(u_3))$. Let $d = a + c - b$. Note that $1 \leq d \leq \min\{a, c\}$. Applying the equalities, we get $\text{pre}_d(\ell(u_1)) = \ell(u_2)[r - a + 1..r - a + d] = \ell(u_3)[c - d + 1..c] = \text{sup}_d(\ell(u_4))$, establishing the existence of the desired overlap. 

Now, consider a $C_4$-free bipartite graph $G = (V_s, V_p, E)$ and let $w$ be a decomposition satisfying the $P_4$-rule. We will prove both Theorem 4.1 and Theorem 4.2 by constructing the following labeling. Let us order the edges $e_1, \ldots, e_{|E|}$ in order of non-decreasing weight. For $0 \leq j \leq |E|$, we define the graph $G^j = (V_s, V_p, \{e_i \in E \mid i \leq j\})$. For a vertex $u$, define $\text{len}_j(u) = \max\{w(e_i) \mid i \leq j, e_i \text{ is incident with } u\}$, if the degree of $u$ in $G^j$ is positive, and 0 otherwise. We will recursively define a labeling $\ell_j$ of $G^j$ such that $|\ell_j(u)| = \text{len}_j(u)$ for all $u$. The initial labeling $\ell_0$ assigns $\epsilon$ to every vertex. Suppose we have a labeling $\ell_j$ for $G^j$, and $e_{j+1} = (u, v)$. Recall that because $w$ satisfies the $P_4$-rule and $G$ is $C_4$-free, $w(u, v) \geq \text{len}_j(u) + \text{len}_j(v) = |\ell_j(u)| + |\ell_j(v)|$. (Note that the inequality holds also in the case when one of the two summands is 0.) Let $A$ be a (possibly empty) string of length $w(u, v) - |\ell_j(u)| - |\ell_j(v)|$ composed of non-repeating characters that do not exist in $\ell_j$. Define $\ell_{j+1}$ as $\ell_{j+1}(x) = \ell_j(x)$ for all $x \notin \{u, v\}$, and $\ell_{j+1}(u) = \ell_{j+1}(v) = \ell_j(v) \cdot A \cdot \ell_j(u)$. We denote the labeling of $G$ as $\ell = \ell_{|E|}$. We will slightly abuse notation in this section, ignoring the fact that a labeling must have labels of the same length. This is inconsequential, because strings can always be padded from the beginning or end with distinct characters without affecting any overlaps.

First, we prove a useful lemma that states that two vertices share a character in the labeling only if they are connected by a path.
Lemma 4.2. Let $c$ be a character that is contained in $\ell_j(u)$ and in $\ell_j(v)$, for some pair of distinct vertices. Then there exists a path between $u$ and $v$ in $G_j$.

Proof. We prove the statement by induction on $m \in \{0, 1, \ldots, |E|\}$. For the base case, $\ell_0$ does not label any positions. Now, assume that $\ell_m$ satisfies the lemma and consider the new positions labeled by $\ell_{m+1}$, with $e_{m+1} = (u, v)$. Recall that $A$ is a possibly empty string of new characters inserted into the middle of the new labels. A position of $u$ labeled with a character from $A$ is adjacent to the position of $v$ labeled with the same character, and since the characters are new, these are the only two positions labeled with this character. Now, each new position of $u$ that is not labeled with a character from $A$ is labeled with a character from $\ell_m(v)$. By the induction hypothesis, $v$ is connected by a path to all vertices with occurrences of the same character in $G^m$, which implies the same statement for $u$ in $G^{m+1}$ (using the fact that $E(G^{m+1}) = E(G^m) \cup \{e_{m+1}\}$). The case of the new characters in the label of $v$ is symmetric. \hfill $\square$

We are now ready to show that $\ell$ achieves $w$ for trees and, if $w$ also satisfies the strict $P_4$-rule, for $C_4$-free graphs.

Lemma 4.3. Let $G$ be a $C_4$-free bipartite graph and let $w$ be a decomposition that satisfies the $P_4$-rule. Then the above defined labeling $\ell$ achieves $w$ if $w$ satisfies the strict $P_4$-rule or if $G$ is acyclic.

Proof. We prove by induction on $j$ that $\ell_j$ achieves $w$ on $G^j$. Suppose that the Lemma holds for $\ell_j$ and consider the effect of adding $e_{j+1} = (u, v)$. Notice that to obtain $\ell_{j+1}$ we only change labels by adding outer characters, hence, any two vertices that overlap by $i$ in $\ell_j$ will also overlap by $i$ in $\ell_{j+1}$. Moreover, only the labels of $u$ and $v$ are changed, and an overlap between $u$ and $v$ of length $w(u, v)$ is created. It remains to show that no shorter overlap is created between $u$ and $v$ and that no new overlap is created involving $u$ or $v$, except the one between $u$ and $v$.

First, consider the case when $w(u, v) > |\ell_j(u)| + |\ell_j(v)|$ and so the middle string ($A$) of the new labels is non-empty. Because the characters of $A$ do not appear in $\ell_j$, we do not create any new overlaps except besides the one between $u$ and $v$ and the only overlap between $u$ and $v$ must be of length $w(u, v)$ since the characters of $A$ must align. Thus $\ell_{j+1}$ achieves $w$ on $G^{j+1}$.

Next, consider the case when $w(u, v) = |\ell_j(u)|$ (the case when $w(u, v) = |\ell_j(v)|$ is symmetric). In this case, $A = \epsilon$, $\ell_j(u) = \epsilon$, and $|\ell_j(v)| > 0$ (since $w(u, v) > 0$). Suppose for the sake of contradiction that there exists a vertex $v' \neq v$ such that $(u, v')$ is not an edge but $inner_k(\ell_{j+1}(u)) =$
inner_k(ℓ_{j+1}(v'))), for some 0 < k ≤ w(u, v). We know, from the construction
of ℓ_j, that there exists a vertex u' such that w(u', v) = |ℓ_j(v)|. We then have
inner_k(ℓ_j(u')) = outer_k(ℓ_j(v)) = inner_k(ℓ_{j+1}(u)) = inner_k(ℓ_{j+1}(v')) =
inner_k(ℓ_j(v')). By the induction hypothesis, there is an edge (u', v') and
w(u', v') ≤ k. The edges (u, v), (v, u'), (u', v') form a P_4, which is also in-
duced because G is C_4-free. Because w(u, v) = w(u', v) ≥ w(u', v') > 0, the
P_4-rule is violated, a contradiction. Therefore no new overlaps are created
involving u. To show that there are no overlaps from u to v smaller than
w(u, v), observe that any such overlap would also be an overlap between u'
and v that is smaller than w(u', v), contradicting the induction hypothesis.
Therefore, ℓ_{j+1} achieves w on G^{j+1}.

It remains to consider the case when w(u, v) = |ℓ_j(u)| + |ℓ_j(v)| and
ℓ_j(u) ≠ ϵ ≠ ℓ_j(v). We first show that this case cannot arise if w satisfies the
strict P_4-rule. There must exist edges in G^j of weights |ℓ_j(u)| and |ℓ_j(v)|
incident with u and v, respectively. These edges, together with (u, v) in
the middle, form a P_4, which must be induced since G does not contain a
C_4. Furthermore, (u, v) achieves the maximum weight. The strict P_4-rule
implies w(u, v) > |ℓ_j(u)| + |ℓ_j(v)|, a contradiction.

Now, assume that G is acyclic, and suppose for the sake of contradiction
that the new labeling creates an overlap between v and a vertex u' ≠ u
(the case of an overlap between u and v' ≠ v is symmetric). Consider the
character c at position |ℓ_j(v)| + 1 of ℓ_{j+1}(v). The length of the overlap be-
tween ℓ_{j+1}(v) and ℓ_{j+1}(u') = ℓ_j(u') must be greater than |ℓ_j(v)|, otherwise
it would have been an overlap in ℓ_j. Thus, ℓ_j(u') must contain c. By con-
struction of v's new label, ℓ_j(u) must also contain c. Applying Lemma 4.2,
there must be a path between u' and u in G^j. On the other hand, the overlap
between v and u' spans (ℓ_j(v))[1], and hence ℓ_j(v) and ℓ_j(u') must share a
character. Applying Lemma 4.2, there must exist a path between u' and v
in G^j. Consequently, there exists a path from u to v in G^j. Combining this
path with e_{j+1} = (u, v), we get a cycle in G^{j+1}, which is a contradiction.

Finally suppose, for the sake of contradiction, that ℓ_{j+1}(u) overlaps
ℓ_{j+1}(v) by some k < w(u, v). By the induction hypothesis, k > |ℓ_j(v)|.
Consider the last character c of ℓ_j(v). It must also appear as the inner
position i = k - |ℓ_j(v)| + 1 in ℓ_{j+1}(u). Since k ≤ w(u, v) - 1, we have
i ≤ w(u, v) - |ℓ_j(v)| = |ℓ_j(u)|, and the i^{th} inner position in ℓ_{j+1}(u) is also
the the i^{th} inner position in ℓ_j(u). Applying Lemma 4.2 to c in ℓ_j(v) and
ℓ_j(u), there must exist a path between u and v in G^j. Combining this path
with e_{j+1} = (u, v), we get a cycle in G^{j+1}, which is a contradiction.

We can now prove Theorems 4.1 and 4.2.

10
Proof of Theorem 4.1. Let \( t = \min \{ k \mid w \text{ is a decomposition of size } k \text{ that satisfies the } P_4\text{-rule} \} \). First, let \( w \) be a decomposition of size \( t \) satisfying the \( P_4\text{-rule} \). Lemma 4.3 states that the above defined labeling \( \ell \) achieves \( u \) and so \( r(T) \leq \max_e(w_e) = t \). For the other direction, consider an overlap labeling \( b \) of \( T \) of minimum length. By Lemma 4.1, the \( b \)-decomposition satisfies the \( P_4\text{-rule} \). Hence, \( r(T) = \text{len}(b) \geq t \). \(\)

Proof of Theorem 4.2. Let \( w \) be a decomposition of size \( t \) satisfying the strict \( P_4\text{-rule} \). By Lemma 4.3, the above defined labeling \( \ell \) achieves \( u \) and so \( r(G) \leq \max_e(w_e) = t \). On the other hand, let \( b \) be an overlap labeling of length \( r(G) \). Define \( w(e) = 2ov\_b(e) - 1 \), for all \( e \in E(G) \). We claim that \( w \) satisfies the strict \( P_4\text{-rule} \), which will imply that \( t \leq \max_e w(e) = 2r(G) - 1 \). To see this, let \( e_1, e_2, e_3 \) be the edges of an arbitrary induced \( P_4 \). Observe that \( w(e_2) = \max\{w(e_1), w(e_2), w(e_3)\} \) if and only if \( ov\_b(e_2) = \max\{ov\_b(e_1), ov\_b(e_2), ov\_b(e_3)\} \). Furthermore, it can be algebraically verified that if \( ov\_b(e_2) \geq ov\_b(e_1) + ov\_b(e_3) \) then \( w(e_2) > w(e_1) + w(e_3) \). By Lemma 4.1, the \( b \)-decomposition satisfies the \( P_4\text{-rule} \) and, therefore, \( w \) satisfies the strict \( P_4\text{-rule} \). \(\)

4.2. General graphs

In the previous subsection, we derived graph theoretic characterizations of readability that are exact for trees and approximate for \( C_4\)-free bipartite graphs. Unfortunately, for a general graph, it is not clear how to construct an overlap labeling from a decomposition satisfying the \( P_4\text{-rule} \) (as we did in Lemma 4.3). In this subsection, we will consider an alternate rule (HUB-rule), which we then use to construct an overlap labeling.

Given \( G = (V_s, V_p, E) \) and a decomposition \( w \) of size \( k \), we define \( G^w_i \), for \( i \in [k] \), as a graph with the same vertices as \( G \) and edges given by \( E(G^w_i) = \{ e \in E \mid w(e) = i \} \). When \( w \) is obvious from the context, we will write \( G_i \) instead of \( G^w_i \). Observe that the edge sets of \( G^w_1, \ldots, G^w_k \) form a partition of \( E \). We say that \( w \) satisfies the hierarchical-union-of-bicliques rule, abbreviated as the HUB-rule, if the following conditions hold: i) for all \( i \in [k] \), \( G^w_i \) is a disjoint union of bicliques, and ii) if two distinct vertices \( u \) and \( v \) are non-isolated twins in \( G^w_i \) for some \( i \in \{2, \ldots, k\} \) then, for all \( j \in [i - 1] \), \( u \) and \( v \) are (possibly isolated) twins in \( G^w_j \). An example of a decomposition satisfying the HUB-rule is any \( w : E \to [k] \) such that \( G^w_i \) is an (arbitrary) disjoint union of bicliques and \( G^w_2, \ldots, G^w_k \) are matchings. We can show that the decomposition implied by any overlap labeling must satisfy the HUB-rule.
Lemma 4.4. Let $\ell$ be an overlap labeling of a bipartite graph $G$. Then the $\ell$-decomposition satisfies the HUB-rule.

Proof. Denote the vertices and edges of the graph as usual: $G = (V_s, V_p, E)$. Consider the $\ell$-decomposition. Fix $i \in [k]$. First, we show that $G_i$ is a union of disjoint bicliques. Observe that a bipartite graph is a disjoint union of bicliques if and only if it contains no induced $P_4$. Therefore, it suffices to prove that $G_i$ does not contain any induced $P_4$. Consider a 4-vertex path $(u, x, y, z)$ in $G_i$. We will show that $G_i$ contains the edge $(u, z)$. Since each edge of the path is in $G_i$, the corresponding overlaps imply that $\text{inner}_i(\ell(u)) = \text{inner}_i(\ell(x)) = \text{inner}_i(\ell(y)) = \text{inner}_i(\ell(z))$. Thus, $\text{inner}_i(\ell(u)) = \text{inner}_i(\ell(z))$. To complete the proof that $(u, z) \in E(G_i)$, it remains to show that $(u, z) \notin E(G_j)$ for all $j \in [i-1]$. For the sake of contradiction suppose $\text{inner}_j(\ell(u)) = \text{inner}_j(\ell(z))$ for some $j \in [i-1]$. Then $\text{inner}_j(\ell(u)) = \text{inner}_j(\ell(x))$ and, consequently, $(u, x)$ is in $E(G_j)$, which contradicts that it is in $E(G_i)$. Therefore, $(u, z) \in E(G_i)$. This completes the proof that $G_i$ is a disjoint union of bicliques.

Next we show that the $\ell$-decomposition is hierarchical, i.e. satisfies the second condition of the HUB-rule definition. Fix $i \in \{2, \ldots, k\}$ and consider two non-isolated twins $u, v$ in $G_i$. By definition of non-isolated twins, there is a vertex $z$ that is adjacent to both $u$ and $v$ in $G_i$. By definition of $G_i$, we get $\text{inner}_i(\ell(u)) = \text{inner}_i(\ell(v)) = \text{inner}_i(\ell(z))$. Therefore, for all $j \in [i-1]$, the corresponding inner affixes of labels of $u$ and $v$ are the same: $\text{inner}_j(\ell(u)) = \text{inner}_j(\ell(v))$. Consequently, in $G_j$, every neighbor of $u$ must be a neighbor of $v$, and vice versa. That is, $u$ and $v$ are twins in $G_j$ for all $j \in [i-1]$, completing the proof of the lemma. \hfill \Box

We define the HUB number of $G$ as the minimum size of a decomposition of $G$ that satisfies the HUB-rule, and denote it by $\text{hub}(G)$. Observe that a decomposition of a graph into matchings (i.e. each $G_i^{\text{m}}$ is a matching) satisfies the HUB-rule. By König’s Line Coloring Theorem, any bipartite graph $G$ can be decomposed into $\Delta(G)$ matchings, where $\Delta(G)$ is the maximum degree of $G$. Thus, $\text{hub}(G) \in [\Delta(G)]$. Clearly, a graph $G$ has $\text{hub}(G) = 1$ if and only if $G$ is a disjoint union of bicliques. The HUB number captures readability in the sense that the readability of a graph family is bounded (by a uniform constant independent of the number of vertices) if and only if its HUB number is bounded. This is captured by the following theorem:

Theorem 4.3. Let $G$ be a bipartite graph. Then

$$\text{hub}(G) \leq r(G) \leq 2^{\text{hub}(G)} - 1.$$
In the remainder of this section, we prove this theorem. The first inequality directly follows from Lemma 4.4 because, by definition of readability, there exists an overlap labeling $\ell$ of length $r(G)$. Then the $\ell$-decomposition of $G$ is of size $r(G)$ and satisfies the HUB-rule, implying $hub(G) \leq r(G)$. To prove the second inequality, we show:

**Lemma 4.5.** Let $w$ be a decomposition of size $k$ of a bipartite graph $G$ that satisfies the HUB-rule. Then there is an overlap labeling of $G$ of length $2^k - 1$.

**Proof.** First, we define a labeling $t$ by applying the following operation due to Braga and Meidanis [BM02]. Given two vertices $u \in V_s$ and $v \in V_p$, a labeling $t$, and a filler character $a$ not used by $t$, the BM operation transforms $t$ by relabeling both $u$ and $v$ with $t(v) \cdot a \cdot t(u)$.

We start by labeling $G_1$ as follows: each biclique $B$ in $G_1$ gets assigned a unique character $a_B$, and each node $v$ in a biclique $B$ gets label $t(v) = a_B$. Next, for $i \in [k-1]$, we iteratively construct a labeling of $G_1 \cup \cdots \cup G_i+1$ from a labeling $t$ of $G_1 \cup \cdots \cup G_i$. We show by induction that the constructed labeling has an additional property that all twins in $G_i+1$ have the same labels and that the length of the labeling is $2^{i+1} - 1$. Observe that the labeling of $G_1$ satisfies this property.

We choose a unique (not previously used) character $a_B$ for each biclique $B$ of $G_i+1$. If $B$ consists of a single vertex $v$, then we assign to $v$ the label $a_B \cdot t(v)$ if $v \in V_s$, and $t(v) \cdot a_B$ if $v \in V_p$. Otherwise, since $w$ satisfied the HUB-rule, all vertices in $B \cap V_s$ are twins in $G_1 \cup \cdots \cup G_i$ and, by the induction hypothesis, are assigned the same labels in $t$. Analogously, $t$ will assign the same labels to all nodes in $B \cap V_p$. Consider an arbitrary edge $(u, v)$ in $B$. We apply the BM operation with character $a_B$ to $(u, v)$ and assign the resulting label $t(v) \cdot a_B \cdot t(u)$ to all nodes in $B$. This completes the construction of labeling of $G_1 \cup \cdots \cup G_i+1$. Observe that it assigns the same labels to all twins in $G_i+1$, and that the length is $2^{i+1} - 1$.

It remains to show that the final labeling is an overlap labeling of $G$. It is easy to see that the initial labeling of $G_1$ is an overlap labeling. Now we show that if $t$ is an overlap labeling of $G_1 \cup \cdots \cup G_i$, our construction yields an overlap labeling of $G_1 \cup \cdots \cup G_i+1$.

Suppose first that $(u, v)$ is an edge of $G_1 \cup \cdots \cup G_i+1$. If $(u, v)$ is an edge of $G_{i+1}$ then, by construction, the labels of $u$ and $v$ after $i + 1$ steps are identical, and consequently they overlap. If $(u, v)$ is not an edge of $G_{i+1}$, then it is an edge of $G_1 \cup \cdots \cup G_i$, and the bicliques $B$ and $B'$ of $G_{i+1}$ containing $u$ and $v$, respectively, are distinct. This implies that the labels of $u$ and $v$ after $i + 1$ steps are of the form $x \cdot a_B \cdot t(u)$ and $t(v) \cdot a_{B'} \cdot y,$
respectively, for some (possibly empty) strings $x, y, a_B$, and $a_{B'}$, where $t(u)$ and $t(v)$ are the respective labels of $u$ and $v$ after $i$ steps. Since, by the induction hypothesis, $t(u)$ and $t(v)$ overlap, so do the extended labels.

Finally, if $(u, v) \in V_s \times V_p$ is a pair of nonadjacent vertices of $G_1 \cup \cdots \cup G_{i+1}$, then $u$ and $v$ are nonadjacent in $G_1 \cup \cdots \cup G_i$. By induction hypothesis, their labels after $i$ steps, $t(u)$ and $t(v)$, do not overlap. Since $u$ and $v$ are also not adjacent in $G_{i+1}$, the bicliques of $G_{i+1}$ containing $u$ and $v$, say $B$ and $B'$, are distinct, and thus the labels of $u$ and $v$ after $i+1$ steps are of the form $x \cdot a_B \cdot t(u)$ and $t(v) \cdot a_{B'} \cdot y$, respectively. Moreover, if both $x \cdot a_B$ and $a_{B'} \cdot y$ are nonempty then $a_B \neq a_{B'}$. Hence, by construction, the two labels do not overlap. This completes the proof. \[ \square \]

This completes the proof of Theorem 4.3, as the second inequality follows directly by choosing a minimum decomposition satisfying the HUB-rule, in which case $k = \text{hub}(G)$.

Note that if $w$ is a decomposition into matchings, then our labeling algorithm behaves identically to the Braga-Meidanis (BM) algorithm [BM02]. However, in the case that $w$ is of size $o(\Delta(G))$, our labeling algorithm gives a better bound than BM. For example, for the $n \times n$ biclique, our algorithm gives a labeling of length 1, while BM gives a labeling of length $2^n - 1$.

5. Lower and upper bounds on readability

In this section, we prove several lower and upper bounds on readability, making use of the characterizations of the previous section.

5.1. Almost all graphs have readability $\Omega(n/\log n)$

In this subsection, we show that, in both the bipartite and directed graph models, there exist graphs with readability at least $\Omega(n/\log n)$, and that in fact almost all graphs have at least this readability.

We will need the following reduction, implicitly shown in [BM02].

**Property 5.1** ([BM02]). Let $G$ be a digraph or a bipartite graph, let $\Sigma$ and $\Sigma'$ be alphabets with $|\Sigma| \geq |\Sigma'| \geq 2$, and let $\ell$ be an overlap labeling of $G$ over $\Sigma$. Then there exists an overlap labeling $\ell'$ of $G$ over $\Sigma'$ such that $\text{len}(\ell') \leq (2 \log |\Sigma'| + 1) \cdot \text{len}(\ell)$.

We can now state and prove the main theorem of the subsection.

**Theorem 5.1.** Almost all graphs in $B_{n \times n}$ (and, respectively, $D_{n}$) have readability $\Omega(n/\log n)$. When restricted to a constant sized alphabet, almost all graphs in $B_{n \times n}$ (and, respectively, $D_{n}$) have readability $\Omega(n)$. 

14
Proof. We prove the lemma by a counting argument. First, consider the case of a constant sized alphabet. Since there are \( n^2 \) pairs of nodes in \([n]^2\) that can form edges in a graph in \( \mathcal{B}_{n \times n} \), the size of \( \mathcal{B}_{n \times n} \) is \( 2^{n^2} \). Let \( a \) be the size of the alphabet. The number of labelings of \( 2n \) nodes with strings of length \( s \) is at most \( a^{2ns} \). In particular, labelings of length \( s = n/(3 \log a) \) can generate no more than \( a^{2n^2/(3 \log a)} = 2^{2n^2/3} \) bipartite graphs, which is in \( o(2^{n^2}) \). Consequently, almost all graphs in \( \mathcal{B}_{n \times n} \) have readability \( \Omega(s) = \Omega(n/\log a) = \Omega(n) \). The proof for \( \mathcal{D}_n \) is analogous and is omitted.

For the variable sized alphabet case, observe that the previous argument shows that only \( o(2^{n^2}) \) graphs in \( \mathcal{B}_{n \times n} \) have readability at most \( n/3 \) over the binary alphabet. It therefore suffices to show that every graph in \( \mathcal{B}_{n \times n} \) of readability at most \( n/(15 \log_2 n) \) (over an unrestricted alphabet) has readability at most \( n/3 \) over the binary alphabet. This is indeed the case. Suppose that \( G \in \mathcal{B}_{n \times n} \) is of readability \( r \leq n/(15 \log_2 n) \), and fix an overlap labeling \( \ell \) of \( G \) of length \( r \). Since \( \ell \) uses \( 2nr \) characters in total, the alphabet size of labeling \( \ell \) can be assumed to be at most \( 2nr \). By Property 5.1, \( G \) has an overlap labeling \( \ell' \) over the binary alphabet such that \( \text{len}(\ell') \leq (2 \log_2(2nr) + 1)r \). Since \( 2nr \leq n^2 \), we have \( 2\log_2(2nr) + 1 \leq 5\log_2 n \) and consequently the readability of \( G \) over the binary alphabet is at most \( \text{len}(\ell') \leq 5r \log_2 n \leq n/3 \). The proof for \( \mathcal{D}_n \) is analogous and is omitted. \( \square \)

5.2. Distinctness and a graph family with readability \( \Omega(n) \)

In this subsection, we give a technique for proving lower bounds and use it to obtain a family of graphs with readability \( \Omega(n) \). For any two vertices \( u \) and \( v \), the distinctness of \( u \) and \( v \) is defined as \( DT(u, v) = \max\{|N(u) \setminus N(v)|, |N(v) \setminus N(u)|\} \). The distinctness of a bipartite graph \( G \), denoted by \( DT(G) \), is defined as the minimum distinctness of any pair of vertices that belong to the same part of the bipartition. The following lemma relates the distinctness and the readability of graphs that are not matchings (for a matching, the readability is 1, provided that it has at least one edge, and 0 otherwise).

Lemma 5.1. For every bipartite graph \( G \) that is not a matching, \( r(G) \geq DT(G) + 1 \).

Proof. By Theorem 4.3, it suffices to show that \( DT(G) \leq \text{hub}(G) - 1 \). Let \( h = \text{hub}(G) \), let \( w : E(G) \to [h] \) be a minimum decomposition of \( G \) satisfying the HUB-rule, and consider the graphs \( G_i = G_i^w \), for \( i \in [h] \). We need to show that \( DT(G) \leq h - 1 \). Suppose first that each \( G_i \) is a matching. Then,
since \( w \) is a decomposition of \( G \), we have \( \Delta(G) \leq h \). Moreover, since \( G \) is not a matching, it has a pair of distinct vertices, say \( u \) and \( v \), with a common neighbor, which implies \( DT(G) \leq DT(u,v) \leq \Delta(G) - 1 \leq h - 1 \).

Suppose now that there exists an index \( j \in [h] \) such that \( G_j \) is not a matching, and let \( j \) be the maximum such index. Then, there exist two distinct vertices in \( G \), say \( u \) and \( v \), that have a common neighbor in \( G_j \), and therefore belong to the same biclique of \( G_j \). It follows that \( u \) and \( v \) are non-isolated twins in \( G_j \). Since \( w \) satisfies the HUB-rule, this implies that \( u \) and \( v \) are twins in each \( G_i \) with \( i \in [j-1] \). Consequently, for each vertex \( x \) in \( G \) adjacent to \( u \) but not to \( v \), the unique \( G_i \) with \( (u,x) \in E(G_i) \) satisfies \( i > j \). By the choice of \( j \), each such \( G_i \) is a matching, and hence there can be at most \( n/j \) such vertices \( x \). Thus \( |N(u) \setminus N(v)| \leq h-j \) and similarly \( |N(v) \setminus N(u)| \leq h-j \), which implies the desired inequality \( DT(G) \leq DT(u,v) \leq h-j \leq h-1 \).

While the distinctness is a much simpler graph parameter than the HUB number, simplicity comes with a price. Namely, the distinctness does not share the nice feature of the HUB number, that of being bounded on exactly the same sets of graphs as the readability. In Section 5.3, we show the existence of graphs (specifically, trees) of distinctness 1 and of arbitrary large readability.

We now introduce a family of graphs, inspired by the Hadamard error correcting code, and apply Lemma 5.1 to show that their readability is at least linear in the number of nodes. We define \( H_k \) as the bipartite graph with vertex sets \( V_s = \{ v_s \mid v \in \{0,1\}^k \setminus \{0^k\} \} \) and \( V_p = \{ v_p \mid v \in \{0,1\}^k \setminus \{0^k\} \} \) and edge set

\[
E(H_k) = \left\{ (v_s,v_p) \in V_s \times V_p \mid \sum_{i=1}^{k} v_s[i]v_p[i] \equiv 1 \pmod{2} \right\}.
\]

In other words, each vertex has a non-zero \( k \)-bit codeword vector associated with it and two vertices are adjacent if the inner product of their codewords is odd. Let \( n = 2^k \). Graph \( H_k \) has \( 2(n-1) \) vertices, all of degree \( n/2 \), and thus \( (n-1)n/2 \) edges. Figure 2 illustrates \( H_3 \). Labelings for \( H_3 \) and \( H_4 \) can be visualized online at http://rchikhi.github.com/readability.

The following lemma shows that every pair of vertices in the same part of the bipartition of \( H_k \) has exactly \( n/4 \) common neighbors, implying that the distinctness of \( H_k \) is \( n/4 \).
Lemma 5.2. In graph $H_k$, if $i$ vertices have a common neighbor, then they have at least $2^{k-i} = n/2^i$ common neighbors. Moreover, if two vertices have a common neighbor, then they have exactly $n/4$ common neighbors.

Proof. Suppose that vertices $w_1, \ldots, w_i \in \{0, 1\}^k \setminus \{0^k\}$ in the same part of the bipartition of $H_k$ have a common neighbor. Then the set $X$ of all vectors $x \in \{0, 1\}^k$ such that $w_j^\top x = \sum_{p=1}^k w_j[p]x[p] \equiv 1 \pmod{2}$ is non-empty. Notice that $X \subseteq \{0, 1\}^k$ is the set of solutions of the equation $Wx = 1$ over the field $GF(2)$, where $W$ is the $i \times k$ matrix with the rows formed by the $w_j$’s, and $1$ is the all-one vector of length $i$. The set $X$ forms an affine subspace of the vector space $\{0, 1\}^k$ over $GF(2)$ of dimension $k - r$, where $r = \text{rank}(W)$. Therefore, vertices $w_1, \ldots, w_i$ have exactly $|X| = 2^{k-r}$ common neighbors. Since $r \leq i$, we obtain $|X| \geq 2^{k-i}$.

If $i = 2$, then the rank of $W$ is exactly 2, which implies the second part of the lemma.

Combining this with Lemma 5.1, we obtain the following theorem.

Theorem 5.2. $r(H_k) \geq n/4 + 1$.

This lower bound also translates to directed graphs: applying Theorem 3.1, there exists digraphs of readability $\Omega(n)$. A major open question is: Do there exist graphs that have exponential readability? We conjecture that they do, and that the graph family $H_k$ has exponential readability. However, since distinctness is $O(n)$, we note that Lemma 5.1 is insufficient for proving stronger than $\Omega(n)$ lower bounds on the readability.
5.3. Trees

The purely graph theoretic characterization of readability given by Theorem 4.1 allows us to derive a sharp upper bound on the readability of trees. The eccentricity of a vertex \( u \) in a connected graph \( G \) is defined as \( \text{ecc}_G(u) = \max_{v \in V(G)} \text{dist}_G(u,v) \), where \( \text{dist}_G(u,v) \) is the number of edges in a shortest path from \( u \) to \( v \). The radius of a graph \( G \) is defined as the minimum eccentricity of a vertex in \( G \), that is \( \text{radius}(G) = \min_{u \in V(G)} \max_{v \in V(G)} \text{dist}_G(u,v) \).

**Theorem 5.3.** For every tree \( T \), \( r(T) \leq \text{radius}(T) \), and this bound is sharp. More precisely, for every \( k \geq 0 \) there exists a tree \( T \) such that \( r(T) = \text{radius}(T) = k \).

**Proof.** Let \( T \) be a tree. If \( T = K_1 \) (the one-vertex tree), then \( \text{radius}(T) = r(T) = 0 \) (note that assigning the empty string to the unique vertex of \( v \) results in an overlap labeling of \( T \)). Now, let \( T \) be of radius \( r \geq 1 \) and let \( v \in V(T) \) be a vertex of \( T \) of minimum eccentricity (that is, \( \text{ecc}_T(v) = r \)). Consider the distance levels of \( T \) from \( v \), that is, \( V_i = \{ w \in V(T) \mid \text{dist}_T(v,w) = i \} \) for \( i \in \{0,1,\ldots,r\} \). Also, for all \( i \in [r] \), let \( E_i \) be the set of edges in \( T \) connecting a vertex in \( V_{i-1} \) with a vertex in \( V_i \). Then \( \{E_1,\ldots,E_r\} \) is a partition of \( E(T) \) and the decomposition \( T \) of \( u \in V(T) \) given by \( w(e) = i \) if and only if \( e \in E_i \) is well defined. We claim that \( w \) satisfies the \( P_4 \)-rule. Let \( P = (v_1,v_2,v_3,v_4) \) be an induced \( P_4 \) in \( T \), and let \( i = w(v_1,v_2), j = w(v_2,v_3), k = w(v_3,v_4) \). Suppose that \( j = \max\{i,j,k\} \). We may assume without loss of generality that \( v_2 \in V_{j-1} \) and \( v_3 \in V_j \). Since \( T \) is a tree, \( v_2 \) is the only neighbor of \( v_3 \) in \( V_{j-1} \), which implies that \( v_4 \in V_{j+1} \) and consequently \( k = j + 1 \), contrary to the assumption \( j = \max\{i,j,k\} \).

Thus, the \( P_4 \)-rule is trivially satisfied for \( u \). By Theorem 4.1, we have \( r(T) \leq \max_{e \in E(T)} w(e) = r = \text{radius}(T) \).

To show that for every \( k \geq 0 \) there exists a tree \( T \) with \( r(T) = \text{radius}(T) = k \), we proceed by induction. We will construct a sequence \( \{(T_i,v_i)\}_{i \geq 0} \) where \( T_i \) is a tree, \( v_i \) is a vertex in \( T_i \) with \( \text{ecc}_{T_i}(v_i) \leq i \), the degree of \( v_i \) in \( T_i \) is \( i \), and \( r(T_i) = \text{radius}(T_i) = i \). For \( i = 0 \), take \( (T_0,v_0) = (K_1,v_0) \) where \( v_0 \) is the unique vertex of \( K_1 \). This clearly has the desired properties. For \( i \geq 1 \), take \( i \) disjoint copies of \( (T_{i-1},v_{i-1}) \), say \( (T_{i-1}^j,v_{i-1}^j) \) for \( j \in [i] \), add a new vertex \( v_i \), and join \( v_i \) by an edge to each \( v_{i-1}^j \) for \( j \in [i] \). Let \( T_i \) be the so constructed tree. Clearly, the degree of \( v_i \) in \( T_i \) is \( i \), and \( \text{ecc}_{T_i}(v_i) \leq 1 + \text{ecc}_{T_i}(v_{i-1}) \leq 1 + (i-1) = i \), which implies that \( \text{radius}(T_i) \leq i \). On the other hand, we will show that \( r(T_i) \geq i \), which together with inequality \( r(T_i) \leq \text{radius}(T_i) \) will imply the desired conclusion.
radius($T_i$) = $r(T_i) = i$. Suppose for a contradiction that $r(T_i) < i$. Then, by Lemma 4.1, there exists a decomposition $w$ of $T_i$ of size $i - 1$ satisfying the $P_4$-rule. In particular, this implies $i \geq 2$. Since the degree of $v_i$ in $T_i$ is $i$, there exist two edges incident with $v_i$, say $(v_i, v_{i-1}^j)$ and $(v_i, v_{i-1}^k)$ for some $j \neq k$ such that $w(v_i, v_{i-1}^j) = w(v_i, v_{i-1}^k)$. Let $w_1$ denote this common value. Let $x$ be a neighbor of $v_{i-1}^j$ in $T_{i-1}^j$. (Note that $x$ exists since $v_{i-1}^j$ is of degree $i - 1 \geq 1$ in $T_{i-1}^j$.) Then, $(x, v_{i-1}^j, v_i, v_{i-1}^k)$ is an induced $P_4$ in $T_i$. We claim that $w(x, v_{i-1}^j) > w_1$. Indeed, if $w(x, v_{i-1}^j) \leq w_1$ then we have $\max\{w(x, v_{i-1}^j), w(v_{i-1}^j, v_i), w(v_i, v_{i-1}^k)\} = \max\{w(x, v_{i-1}^j), w_1, w_1\} = w_1$, while $w_1 \neq w_1 + w(x, v_{i-1}^j)$, contrary to the $P_4$-rule. Since $x$ was an arbitrary neighbor of $v_{i-1}^j$ in $T_{i-1}^j$, we infer that every edge $e$ in $T_{i-1}^j$ incident with $v_{i-1}^j$ satisfies $w(e) > w_1$. In particular, this leaves a set of at most $i - 2$ different values that can appear on these $i - 1$ edges (the value $w_1$ is excluded), and hence again there must be two edges of the same weight, say $w_2$. Clearly, $w_2 > w_1$ and $i > 2$. Proceeding inductively, we construct a sequence of edges $e_1, e_2, \ldots, e_t$ forming a path in $T_i$ from $v_i$ to a leaf and satisfying $w_1 < w_2 < \ldots < w_t$, where $w_i = w(e_i)$. This implies that all the weights $w_1, \ldots, w_t$ are distinct, contrary to the fact that the range of $w$ is contained in the set $[i - 1]$. This contradiction shows that $r(T_i) \geq i$ and completes the proof. 

Note that for every $k \geq 2$, the tree $T_k$ of radius $k$ constructed in the proof of Theorem 4.1 has a pair of leaves in the same part of the bipartition and is therefore of distinctness 1. This shows that the readability of a graph cannot be upper-bounded by any function of its distinctness (cf. Lemma 5.1).

6. Conclusion

In this paper, we define a graph parameter called readability and initiate a study of its asymptotic behavior. We give purely graph theoretic parameters (i.e., without reference to strings) that are exactly (respectively, asymptotically) equivalent to readability for trees (respectively, $C_4$-free graphs); however, for general graphs, the HUB number is equivalent to readability only in the sense that it is bounded on the same set of graphs. While an $\ell$-decomposition always satisfies the HUB-rule, the converse is not true. For example, a decomposition of $P_4$ with weights $4, 5, 3$ satisfies the HUB-rule but cannot be achieved by an overlap labeling (by Lemma 4.1). For this reason, the upper bound given by Lemma 4.5 leaves a gap with the lower
bound of Lemma 4.4. We are able to describe other properties that an \( \ell \)-decomposition must satisfy (not included in the paper), however, we are not able to exploit them to close the gap. It is a very interesting direction to find other necessary rules that would lead to a graph theoretic parameter that would more tightly match readability on general graphs than the HUB number.

Consider \( r(n) = \max \{ r(D) \mid D \text{ is a digraph on } n \text{ vertices} \} \). We have shown \( r(n) = \Omega(n) \) and know from [BM02] that \( r(n) = O(2^n) \). Can this gap be closed? Do there exist graphs with readability \( \Theta(2^n) \) (as we conjecture), or, for example, is readability always bounded by a polynomial in \( n \)? Questions regarding complexity are also unexplored, e.g., given a digraph, is it NP-hard to compute its readability? For applications to bioinformatics, the length of reads can be said to be poly-logarithmic in the number of vertices. It would thus be interesting to further study the structure of graphs that have poly-logarithmic readability.

Acknowledgements. P.M. and M.M. would like to thank Marcin Kamiński for preliminary discussions. P.M. was supported in part by NSF awards DBI-1356529 and CAREER award IIS-1453527. M.M. was supported in part by the Slovenian Research Agency (I0-0035, research program P1-0285 and research projects N1-0032, J1-5433, J1-6720, and J1-6743). S.R. was supported in part by NSF CAREER award CCF-0845701, NSF award AF-1422975, and and Boston University’s Hariri Institute for Computing and Center for Reliable Information Systems and Cyber Security.

References

[BBT13] Guy Bresler, Ma’ayan Bresler, and David Tse. Optimal assembly for high throughput shotgun sequencing. *BMC Bioinformatics*, 14(Suppl 5):S18, 2013.

[BFK+02] Jacek Blazewicz, Piotr Formanowicz, Marta Kasprzak, Petra Schuurman, and Gerhard J. Woeginger. DNA sequencing, Eulerian graphs, and the exact perfect matching problem. In *Graph-Theoretic Concepts in Computer Science*, pages 13–24. Springer, 2002.

[BFKK02] Jacek Blazewicz, Piotr Formanowicz, Marta Kasprzak, and Daniel Kobler. On the recognition of de Bruijn graphs and their induced subgraphs. *Discrete Mathematics*, 245(1):81–92, 2002.
[BHKdW99] Jacek Blazewicz, Alain Hertz, Daniel Kobler, and Dominique de Werra. On some properties of DNA graphs. *Discrete Applied Mathematics*, 98(1):1–19, 1999.

[BM02] Marília D. V. Braga and Joao Meidanis. An algorithm that builds a set of strings given its overlap graph. In *LATIN 2002: Theoretical Informatics*, 5th Latin American Symposium, Cancun, Mexico, April 3-6, 2002, Proceedings, pages 52–63, 2002.

[BM08] John A. Bondy and Uppaluri S. R. Murty. *Graph Theory*, volume 244 of *Graduate Texts in Mathematics*. Springer, New York, 2008.

[GP14] Theodoros P. Gevezes and Leonidas S. Pitsoulis. Recognition of overlap graphs. *Journal of Combinatorial Optimization*, 28(1):25–37, 2014.

[IW95] Ramana M. Idury and Michael S. Waterman. A new algorithm for DNA sequence assembly. *Journal of Computational Biology*, 2(2):291–306, 1995.

[LZ07] Xianyue Li and Heping Zhang. Characterizations for some types of DNA graphs. *Journal of Mathematical Chemistry*, 42(1):65–79, 2007.

[LZ10] Xianyue Li and Heping Zhang. Embedding on alphabet overlap digraphs. *Journal of Mathematical Chemistry*, 47(1):62–71, 2010.

[MGMB07] Paul Medvedev, Konstantinos Georgiou, Gene Myers, and Michael Brudno. Computability of models for sequence assembly. In *Algorithms in Bioinformatics*, pages 289–301. Springer, 2007.

[MKS10] Jason R Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010.

[Mye05] Eugene W. Myers. The fragment assembly string graph. In *ECCB/JBI*, page 85, 2005.

[NP09] Niranjan Nagarajan and Mihai Pop. Parametric complexity of sequence assembly: theory and applications to next generation
sequencing. *Journal of computational biology*, 16(7):897–908, 2009.

[NP13] Niranjan Nagarajan and Mihai Pop. Sequence assembly demystified. *Nature Reviews Genetics*, 14(3):157–167, 2013.

[PSW03] Rudi Pendavingh, Petra Schuurman, and Gerhard J. Woeginger. Recognizing DNA graphs is difficult. *Discrete Applied Mathematics*, 127(1):85–94, 2003.

[Swe00] Z Sweedyk. A $2\frac{1}{2}$-approximation algorithm for Shortest Superstring. *SIAM Journal on Computing*, 29(3):954–986, 2000.

[TU88] Jorma Tarhio and Esko Ukkonen. A greedy approximation algorithm for constructing shortest common superstrings. *Theoretical Computer Science*, 57(1):131–145, 1988.