Bipartite Graphs of Small Readability[☆]

Rayan Chikhi

CNRS, UMR 9189, France.

Vladan Jovičić

ENS Lyon, France.

Stefan Kratsch

Institut für Informatik, Humboldt-Universität zu Berlin, Germany.

Paul Medvedev

The Pennsylvania State University, USA.

Martin Milanič

IAM and FAMNIT, University of Primorska, Koper, Slovenia.

Sofya Raskhodnikova, Nithin Varma

Boston University, USA.

Abstract

We study a parameter of bipartite graphs called readability, introduced by Chikhi et al. (*Discrete Applied Mathematics*, 2016) and motivated by applications of overlap graphs in bioinformatics. The behavior of the parameter is poorly understood. The complexity of computing it is open and it is not known whether the decision version of the problem is in NP. The only known upper bound on the readability of a bipartite graph (following from a work of Braga and Meidanis, *LATIN* 2002) is exponential in the maximum degree of the graph.

Graphs that arise in bioinformatics applications have low readability. In this paper, we focus on graph families with readability $o(n)$, where n is the number of vertices. We show that the readability of n-vertex bipartite chain graphs is between $\Omega(\log n)$ and $\mathcal{O}(\sqrt{n})$. We give an efficiently testable characterization of bipartite graphs of readability at most 2 and completely determine the readability of grids, showing in particular that their readability never exceeds 3. As a consequence, we obtain a polynomial time algorithm to determine the readability of induced subgraphs of grids. One of the highlights of our techniques is the appearance of Euler's totient function in the analysis of the readability of bipartite chain graphs. We also develop a new technique for proving lower bounds on readability, which is applicable to dense graphs with a large number of distinct degrees.

Keywords: readability, bipartite graph, grid graph, Euler's totient function

[☆]A preliminary version of this work appeared in the proceedings of COCOON 2018 [6].

Email addresses: `rayan.chikhi@univ-lille1.fr` (Rayan Chikhi), `vladan94.jovicic@gmail.com` (Vladan Jovičić), `kratsch@informatik.hu-berlin.de` (Stefan Kratsch), `pashadag@cse.psu.edu` (Paul Medvedev), `martin.milanic@upr.si` (Martin Milanič), `sofya@bu.edu` (Sofya Raskhodnikova), `nvarma@bu.edu` (Nithin Varma)

1. Introduction

In this work, we further the study of *readability* of bipartite graphs initiated by Chikhi et al. [7]. Given a bipartite graph $G = (V_s, V_p, E)$, an *overlap labeling* of G is a mapping from vertices to strings, called labels, such that for all $u \in V_s$ and $v \in V_p$ there is an edge between u and v if and only if the label of u *overlaps* with the label of v (i.e., a non-empty suffix of u's label is equal to a prefix of v's label). The *length* of an overlap labeling of G is the maximum length (i.e., number of characters) of a label. The *readability* of G, denoted $r(G)$, is the smallest nonnegative integer r such that there is an overlap labeling of G of length r. We emphasize that in this definition, no restriction is placed on the alphabet. One could also consider variants of readability parameterized by the size of the alphabet. A result of Braga and Meidanis [5] implies that these variants are within constant factors of each other, where the constants are logarithmic in the alphabet sizes.

Initially, the notion of readability arose in the study of overlap digraphs. Given a set S of strings, the *overlap digraph* of S is the digraph with vertex set S, with an edge from $s \in S$ to $s' \in S$ if and only if s overlaps s'. Overlap digraphs constructed from DNA strings have various applications in bioinformatics. In particular, in the context of genome assembly, variants of overlap digraphs appear as either de Bruijn graphs [12] or string graphs [19, 22] and are the foundation of most modern assemblers (see [18, 20] for a survey). Several graph-theoretic parameters of overlap digraphs have been studied [3, 2, 4, 10, 16, 17, 21, 24], with a nice survey in [15]. As shown by Braga and Meidanis [5], every digraph without multiple edges but with loops allowed is the overlap digraph of some set of strings. Therefore, one can define the readability of a digraph D as the smallest maximum length of a set of strings the overlap digraph of which is isomorphic to D.

Chikhi et al. showed in [7, Theorem 3.1] that there is a bijection ϕ from the set of all n-vertex digraphs to the set of all bipartite graphs with n vertices in each part such that for every n-vertex digraph D with at least one edge, the readability r of D and the readability r' of its image $\phi(D)$ (in the bipartite sense, as defined above), are related by the inequalities $r' < r \leq 2 \cdot r' - 1$. Therefore, the readability of digraphs is asymptotically equivalent to that of balanced bipartite graphs[1], up to (roughly) a factor of 2. This relation between the readability of a digraph and the corresponding bipartite graph, along with the fact that overlap digraphs of genomes typically have low readability, motivate the study of bipartite graphs with low readability. In this work we derive several results about bipartite graphs with readability sublinear in the number of vertices.

For general bipartite graphs, the only known upper bound on readability is implicit from the work of Braga and Meidanis on overlap digraphs [5]. As observed by Chikhi et al. [7, Theorem 4.3], it follows from the construction in [5] that the readability of a bipartite graph is well defined and at most $2^{\Delta+1} - 1$, where Δ is the maximum degree of the graph. Moreover, Chikhi et al. [7, Theorem 5.1] showed that a $1 - o(1)$ fraction of bipartite graphs with n vertices in each part have readability $\Omega(n/\log n)$. They also constructed [7, Theorem 5.2] an explicit graph family (called Hadamard graphs) with readability $\Omega(n)$.

For trees, readability can be defined in terms of an integer function on the edges, without any reference to strings or their overlaps [7, Theorem 4.1]. In this work, we reveal another connection to number theory, through Euler's totient function, and use it to prove an upper bound on the readability of bipartite chain graphs.

So far, our understanding of readability has been hindered by the difficulty of proving lower bounds. Chikhi et al. [7] developed a lower bound technique that is applicable to graphs with sufficiently different neighborhoods for any two vertices in each part. In this work, we add another technique to the toolbox. Our technique is applicable to dense graphs with a large number of distinct degrees. We apply this technique to obtain a lower bound on readability of bipartite chain graphs.

We give a characterization of bipartite graphs of readability at most 2 and use this characterization to obtain a polynomial time algorithm for checking if a graph has readability at most 2. This is the first nontrivial result of this kind: graphs of readability at most 1 are extremely simple (disjoint unions of complete bipartite graphs, see [7]), whereas the problem of recognizing graphs of readability 3 is open.

[1] A bipartite graph $G = (V_s, V_p, E)$ is said to be *balanced* if $|V_s| = |V_p|$.

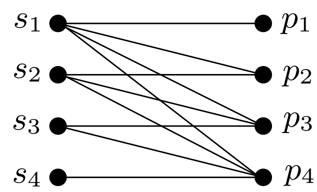


Figure 1: The graph $C_{4,4}$

We also give a formula for the readability of grids, showing in particular that their readability is at most 3. As a corollary, we obtain a polynomial time algorithm to determine the readability of induced subgraphs of grids.

1.1. Our Results and Structure of the Paper

Preliminaries are summarized in Section 2; here we only state some of the most important technical facts. In the study of readability, it suffices to consider bipartite graphs that are connected and twin-free. A bipartite graph is *twin-free* if no two vertices in the same part have the same set of neighbors [7]. Since connected bipartite graphs have a unique bipartition up to swapping the two parts, some of our results are stated without specifying the bipartition.

Bounds on the readability of bipartite chain graphs (Section 3). Bipartite chain graphs are the bipartite analogue of a family of digraphs that occur naturally as subgraphs of overlap graphs of genomes. A *bipartite chain graph* is a bipartite graph $G = (V_s, V_p, E)$ such that the vertices in V_s (or V_p) can be linearly ordered with respect to inclusion of their neighborhoods. That is, we can write $V_s = \{v_1, \ldots, v_k\}$ so that $N(v_1) \subseteq \ldots \subseteq N(v_k)$ (where $N(u)$ denotes the set of u's neighbors). A bipartite chain graph that is connected and twin-free must have the same number of vertices on either side. For each $n \in \mathbb{N}$, there is, up to isomorphism, a unique connected twin-free bipartite chain graph with n vertices in each part, denoted $C_{n,n}$. The graph $C_{n,n}$ is (V_s, V_p, E) where $V_s = \{s_1, \ldots, s_n\}, V_p = \{p_1, \ldots, p_n\}$, and $E = \{(s_i, p_j) \mid 1 \leq i \leq j \leq n\}$. The graph $C_{4,4}$ is shown in Figure 1. We prove an upper and a lower bound on the readability of $C_{n,n}$.

Theorem 1. *For all $n \in \mathbb{N}$, the graph $C_{n,n}$ has readability $\mathcal{O}(\sqrt{n})$, with labels over an alphabet of size 3.*

We prove Theorem 1 by giving an efficient algorithm that constructs an overlap labeling of $C_{n,n}$ of length $\mathcal{O}(\sqrt{n})$ using strings over an alphabet of size 3.

Theorem 2. *For all $n \in \mathbb{N}$, the graph $C_{n,n}$ has readability $\Omega(\log n)$.*

Characterization of bipartite graphs with readability at most 2 (Section 4). Let C_t for $t \in \mathbb{N}$ denote the cycle with t vertices. The *domino* is the graph obtained from the cycle C_6 by adding an edge between a pair of two diametrically opposite vertices (see Figure 4 on p. 10). For a graph G and a set $U \subseteq V(G)$, let $G[U]$ denote the subgraph of G induced by U.

Chikhi et al. [7] proved that every bipartite graph with readability at most 1 is a disjoint union of complete bipartite graphs (also called bicliques). The characterization in the following theorem extends our understanding to graphs of readability at most 2. Recall that a *matching* in a graph is a set of pairwise disjoint edges.

Theorem 3. *A twin-free bipartite graph G has readability at most 2 if and only if G has a matching M such that the graph $G' = G - M$ satisfies the following properties:*

1. *G' is a disjoint union of complete bipartite graphs.*
2. *For $U \subseteq V(G)$, if $G[U]$ is a C_6, then $G'[U]$ is the disjoint union of three edges.*
3. *For $U \subseteq V(G)$, if $G[U]$ is a domino, then $G'[U]$ is the disjoint union of a C_4 and an edge.*

Note that Theorem 3 expresses a condition on vertex labels of a bipartite graph in purely graph-theoretic terms. This reduces the problem of deciding if a graph has readability at most 2 to checking the existence of a matching with a specific property.

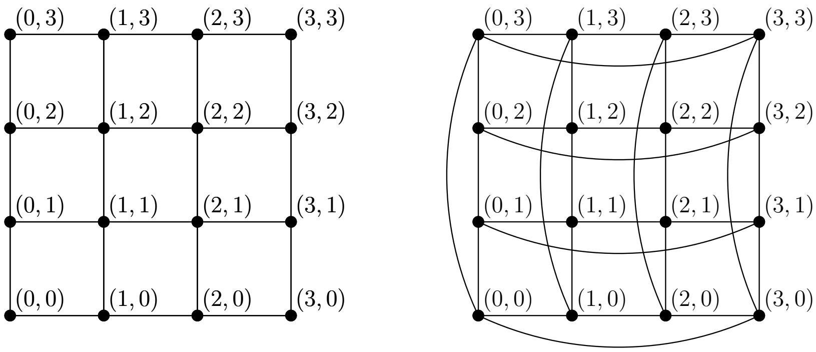


Figure 2: The 4×4 grid $G_{4,4}$ and toroidal grid $TG_{4,4}$.

An efficient algorithm for readability 2 (Section 5). It is unknown whether computing the readability of a given bipartite graph is NP-hard. In fact, it is not even known whether the decision version of the problem is in NP, as the only upper bound on the readability of a bipartite graph with n vertices in each part is $\mathcal{O}(2^n)$ [5]. We make progress on this front by showing that for readability 2, the decision version is polynomial time solvable.

Theorem 4. *There exists an algorithm that, given a bipartite graph G, decides in polynomial time whether G has readability at most 2.*

Moreover, if the answer is "yes", the algorithm can also produce an overlap labeling of length at most 2.

Readability of grids and grid graphs (Section 6). We give a full characterization of the readability of grids. A *(two-dimensional) grid* is a graph $G_{m,n}$ with vertex set $\{0, 1, \dots, m-1\} \times \{0, 1, \dots, n-1\}$ such that there is an edge between two vertices if and only if the L_1-distance between them is 1. An example is shown in Figure 2. The following theorem fully settles the question of readability of grids.

Theorem 5. *For any two positive integers m, n with $m \leq n$, we have*

$$r(G_{m,n}) = \begin{cases} 3, & \text{if } m \geq 3; \\ 2, & \text{if } (m = 2 \text{ and } n \geq 3) \text{ or } (m = 1 \text{ and } n \geq 4); \\ 1, & \text{if } (m, n) \in \{(1,2), (1,3), (2,2)\}; \\ 0, & \text{if } m = n = 1. \end{cases}$$

Theorem 5 has an algorithmic implication for the readability of grid graphs, where a *grid graph* is an induced subgraph of a grid. Several problems are known to be NP-hard on the class of grid graphs, including Hamiltonicity problems [13], various layout problems [9], and others (see, e.g., [8]). We show that unless P = NP, this is not the case for the readability problem.

Corollary 1. *The readability of a given grid graph can be computed in polynomial time.*

1.2. Technical Overview

We now give a brief description of our techniques. The key to proving the upper bound on the readability of bipartite chain graphs is understanding the combinatorics of the following process. We start with the sequence $(1, 2)$. The process consists of a series of rounds, and as a convention, we start at round 3: we write $3\ (= 1 + 2)$ between 1 and 2 and obtain the sequence $(1, 3, 2)$. More generally, in round r, we insert r between all the consecutive pairs of numbers in the current sequence that sum up to r. Thus, we obtain $(1, 4, 3, 2)$ in round 4, then $(1, 5, 4, 3, 5, 2)$ in round 5, and so on. The question is to determine the length of the sequence formed in round r as a function of r. We prove that this length is $\frac{1}{2} \sum_{k=1}^{r} \varphi(k) = \Theta(r^2)$, where $\varphi(k)$ is the famous Euler's totient function denoting the number of integers in $\{1, \dots, k\}$ that are coprime to k.

To prove our lower bound on the readability of bipartite chain graphs, we define a special sequence of subgraphs of the bipartite chain graph such that the number of graphs in the sequence is a lower bound on the readability. The sequence that we define has the additional property that if two vertices in the same part have the same set of neighbors in one of the graphs, then they have the same set of neighbors in all of the preceding graphs in the sequence. If the readability is very small, then we cannot simultaneously cover all the edges incident with two large-degree nodes as well as have their degrees distinct. The only properties of the connected twin-free bipartite chain graph that our proof uses are that it is dense and all vertices in the same part have distinct degrees. Hence, this technique is more broadly applicable to any class of dense graphs with a large number of distinct degrees.

Our characterization of graphs of readability at most 2, roughly speaking, states that a twin-free bipartite graph has readability at most 2 if and only if the graph can be decomposed into two subgraphs G_1 and G_2 such that G_1 is a (vertex-)disjoint union of bicliques and G_2 is a matching satisfying some additional properties. For $i \in \{1, 2\}$, the edges in G_i model overlaps of length exactly i. The heart of the proof lies in observing that for each pair of bicliques in the first subgraph, there can be at most one matching edge in the second subgraph that has its left endpoint in the first biclique and the right endpoint in the second biclique.

To derive a polynomial time algorithm for recognizing graphs of readability two, we first reduce the problem to connected twin-free graphs of maximum degree at least three. For such graphs, we show that the constraints from our characterization of graphs of readability at most 2 can be expressed with a 2SAT formula having variables on edges and modeling the selection of edges forming a matching to form the graph G_2 of the decomposition.

In order to determine the readability of grids, we establish upper and lower bounds and in both cases use the fact that readability is monotone under induced subgraphs (that is, the readability of a graph is at least the readability of each of its induced subgraphs). The upper bound is derived by observing that every grid is an induced subgraph of some $4n \times 4n$ toroidal grid (see Figure 2) and exploiting the symmetric structure of such toroidal grids to show that their readability is at most 3. This is the most interesting part of our proof and involves partitioning the edges of a $4n \times 4n$ toroidal grid into three sets and coming up with labels of length at most 3 for each vertex based on the containment of the four edges incident with the vertex in each of these three parts. Our characterization of graphs of readability at most 2 is a helpful ingredient in proving the lower bound on the readability of grids, where we construct a small subgraph of the grid for which our characterization easily implies that its readability is at least 3.

2. Preliminaries

For a string x, let $\text{pre}_i(x)$ (respectively, $\text{suf}_i(x)$) denote the prefix (respectively, suffix) of x of length i. A string x *overlaps* another string y if there exists an i with $1 \leq i \leq \min\{|x|, |y|\}$ such that $\text{suf}_i(x) = \text{pre}_i(y)$. If $1 \leq i < \min\{|x|, |y|\}$, we say that *x properly overlaps* with y. For a positive integer k, we denote by $[k]$ the set $\{1, \ldots, k\}$. Let $G = (V, E)$ be a (finite, simple, undirected) graph. If G is a connected bipartite graph, then it has a unique bipartition (up to the order of the parts). In this paper, we consider bipartite graphs $G = (V, E)$. If the bipartition $V = V_s \cup V_p$ is specified, we denote such graphs by $G = (V_s, V_p, E)$. Edges of a bipartite graph G are denoted by $\{u, v\}$ or by (u, v) (which implicitly implies that $u \in V_s$ and $v \in V_p$). We respect bipartitions when we perform graph operations such as taking an induced subgraph and disjoint union. For example, we say that a bipartite graph $G_1 = (V_s^1, V_p^1, E_1)$ is an *induced subgraph* of a bipartite graph $G_2 = (V_s^2, V_p^2, E_2)$ if $V_s^1 \subseteq V_s^2$, $V_p^1 \subseteq V_p^2$, and $E_1 = E_2 \cap \{(x, y) : x \in V_s^1, y \in V_p^1\}$. The *disjoint union* of two vertex-disjoint bipartite graphs $G_1 = (V_s^1, V_p^1, E_1)$ and $G_2 = (V_s^2, V_p^2, E_2)$ is the bipartite graph $(V_s^1 \cup V_s^2, V_p^1 \cup V_p^2, E_1 \cup E_2)$.

The path on n vertices is denoted by P_n. Given two graphs F and G, graph G is said to be *F-free* if no induced subgraph of G is isomorphic to F. Two vertices u, v in a bipartite graph are called *twins* if they belong to the same part of the bipartition and have the same neighbors (that is, if $N(u) = N(v)$). Given a bipartite graph $G = (V_s, V_p, E)$ we can define its *twin-free reduction* $TF(G)$ as the graph with vertices being the equivalence classes of the twin relation on $V(G)$ (that is, $x \sim y$ if and only if x and y are twins in G), and two classes X and Y are adjacent if and only if $(x, y) \in E$ for some $x \in X$ and $y \in Y$. For graph-theoretic terms not defined here, we refer to [25].

We now state some basic results for later use.

Lemma 1. *Let G and H be two bipartite graphs. Then:*

(a) If G is an induced subgraph of H, then $r(G) \leq r(H)$.
(b) If F is the disjoint union of G and H, then $r(F) = \max\{r(G), r(H)\}$.
(c) The readability of G is the same for all bipartitions of $V(G)$.
(d) $r(G) = r(TF(G))$.

Proof. (a) If ℓ is any overlap labeling for H then the restriction of ℓ to $V(G)$ yields an overlap labeling for G. Thus, $r(G) \leq r(H)$.

(b) Part (a) implies that $r(G) \leq r(F)$ and $r(H) \leq r(F)$; thus $r(F) \geq \max\{r(G), r(H)\}$. On the other hand, let ℓ_G and ℓ_H be optimal labelings of G and H, over Σ_G and Σ_H, respectively. By introducing new characters if necessary, we may assume that $\Sigma_G \cap \Sigma_H = \emptyset$. Thus, the combined labeling ℓ of F over $\Sigma = \Sigma_G \cup \Sigma_H$, defined as

$$\ell(x) = \begin{cases} \ell_G(x), & \text{if } x \in V(G); \\ \ell_H(x), & \text{if } x \in V(H). \end{cases}$$

for all $x \in V(F)$, is an overlap labeling of F, showing that $r(F) \leq \max\{r(G), r(H)\}$.

(c) By part (b), the readability of G is the maximum readability of a connected component of G. Therefore, it is sufficient to prove the lemma for the case when G is connected. Every connected graph has a unique bipartition, up to switching the roles of V_s and V_t. Switching the roles of V_s and V_t in a graph does not affect its readability, because an overlap labeling of the new graph can be obtained by reversing all the labels in the overlap labeling of the original graph. Thus, the readability of G is not affected by the choice of bipartition of $V(G)$.

(d) It suffices to prove that for a pair of twins u and v, $r(G) = r(G - u)$. By part (a), we have $r(G - u) \leq r(G)$. Conversely, an optimal overlap labeling ℓ of $G - u$ can be extended to an overlap labeling ℓ' of G of the same maximum length as ℓ by setting, for all $x \in V(G)$,

$$\ell'(x) = \begin{cases} \ell(x), & \text{if } x \in V(G) \setminus \{v\}; \\ \ell(u), & \text{if } x = v. \end{cases}$$

Thus, $r(G) \leq r(G - u)$. □

Lemma 1(b) shows that the study of readability reduces to the case of connected bipartite graphs. By Lemma 1(c), the readability of a bipartite graph is well defined even if a bipartition is not given in advance. We state our results without specifying a bipartition in Sections 4-5. Lemma 1(d) further shows that to understand the readability of connected bipartite graphs, it suffices to study the readability of connected twin-free bipartite graphs.

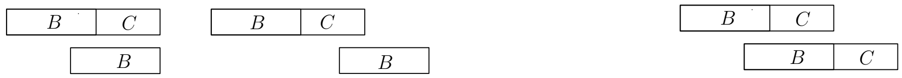
3. Readability of Bipartite Chain Graphs

In this section, we prove an upper bound (Section 3.1) and a lower bound (Section 3.2) on the readability of twin-free bipartite chain graphs $C_{n,n}$. Recall that the graph $C_{n,n}$ is (V_s, V_p, E) where $V_s = \{s_1, \ldots, s_n\}$, $V_p = \{p_1, \ldots, p_n\}$, and $E = \{(s_i, p_j) \mid 1 \leq i \leq j \leq n\}$.

3.1. Upper Bound

Theorem 1. *For all $n \in \mathbb{N}$, the graph $C_{n,n}$ has readability $\mathcal{O}(\sqrt{n})$, with labels over an alphabet of size 3.*

To prove Theorem 1, we construct a labeling ℓ of length $\mathcal{O}(\sqrt{n})$ for $C_{n,n}$ that satisfies (1) $\ell(s_i) = \ell(p_i)$ for all $i \in [n]$, and (2) $\ell(s_i)$ properly overlaps $\ell(s_j)$ if and only if $i < j$. It is easy to see that such an ℓ will be a valid overlap labeling of $C_{n,n}$. As the labels on either side of the bipartition are equal, we will just come up with a sequence of n strings to be assigned to one of the sides of $C_{n,n}$ such that the strings satisfy condition (2) above.



(a) BC does not overlap B.

(b) BC has no proper overlap with itself.

Figure 3: Overlaps in the proof of Lemma 2

Definition 1. A sequence of strings $(s_1, \ldots, s_t)$ is *forward-matching* if

- $\forall i \in [t]$, string s_i does not have a proper overlap with itself and
- $\forall i, j \in [t]$, string s_i overlaps string s_j if and only if $i \leq j$.

Given an integer $r \geq 2$, we will show how to construct a forward-matching sequence S_r with $\Theta(r^2)$ strings, each of length at most r, over an alphabet of size 3. This will imply an overlap labeling of length $\mathcal{O}(\sqrt{n})$ for $C_{n,n}$, proving Theorem 1. The following lemma is crucial for this construction.

Lemma 2. *For all integers $t \geq 2$ and all $i \in [t-1]$, if $(s_1, \ldots, s_t)$ is forward-matching, then so is $(s_1, \ldots, s_i, s_i s_{i+1}, s_{i+1}, \ldots, s_t)$.*

Proof. For the purposes of notation, let A be an arbitrary string from $s_1, \ldots, s_{i-1}$ (if it exists), let $B = s_i$, $C = s_{i+1}$, and let D be an arbitrary string from $s_{i+2}, \ldots, s_t$ (if it exists). The reader can easily verify that A and B overlap with the new string BC, and BC overlaps with C and D, as desired. What remains to show is that there are no undesired overlaps. Suppose for the sake of contradiction that BC overlaps B, and let i be the length of any such overlap. If $\mathrm{suf}_i(BC)$ only includes characters from C, then C overlaps B; if it includes characters from B (and the entire C) then B has a proper overlap with itself (see Figure 3a). In either case, we reach a contradiction. So, BC does not overlap B. By a symmetric argument, C does not overlap BC.

Next, suppose for the sake of contradiction that BC overlaps A, and let i be the length of any such overlap. If $\mathrm{suf}_i(BC)$ only includes characters from C, then C overlaps A; if it includes characters from B (and the entire C) then B overlaps A. In either case, we reach a contradiction. So, BC does not overlap A. By a symmetric argument, D does not overlap BC.

Finally, suppose for the sake of contradiction that BC has a proper overlap with itself, and let i be the length of any such overlap. Since C does not overlap BC, it follows that $\mathrm{suf}_i(BC)$ must include characters from B and the entire C. But then B has a proper overlap with B, a contradiction (see Figure 3b). So, BC does not have a proper overlap with itself, completing the proof. $\square$

Now, we show how to construct a forward-matching sequence S_r. For the base case, we let $S_2 = (20, 0, 01)$. It can be easily verified that S_2 is forward-matching. Inductively, let S_r for $r > 2$ denote the sequence obtained from S_{r-1} by applying the operation in Lemma 2 to all indices i such that $s_i s_{i+1}$ is of length r, that is, add all obtainable strings of length r. Let B_r, for all integers $r \geq 2$, be the sequence of lengths of strings in S_r. We can obtain B_r directly from B_{r-1} by performing the following operation: for each consecutive pair of numbers x, y in B_{r-1}, if $x + y = r$ then insert r between x and y. Note that there is a mirror symmetry to the sequences with respect to the middle element, 1. The right sides of the first 6 sequences B_r, starting from the middle element, are as follows:

$r=2$	1	2								
$r=3$	1	3	2							
$r=4$	1	4	3	2						
$r=5$	1	5	4	3	5	2				
$r=6$	1	6	5	4	3	5	2			
$r=7$	1	7	6	5	4	7	3	5	7	2

It turns out that $|B_r|$, and, by extension, $|S_r|$, is closely related to the totient summatory function [23], also called the partial sums of Euler's totient function. This is the function $\Phi(r) = \sum_{k=1}^{r} \varphi(k)$, where $\varphi(k)$

is the number of integers in $[k]$ that are coprime to k. The asymptotic behavior of $\Phi(r)$ is well known: $\Phi(n) = \frac{3n^2}{\pi^2} + \mathcal{O}(n \log n)$ [11, p. 268]. The following lemma therefore implies $|S_r| = |B_r| = \Theta(r^2)$, completing the proof of Theorem 1.

Lemma 3. *For all integers $r \geq 2$, the length of the sequence B_r is $\Phi(r) + 1$.*

Proof. For the base case, observe that $|B_2| = 3 = \Phi(2) + 1$. In general, consider the case of $r \geq 3$.

Definition 2. Two elements of B_r are called *neighbors in B_r* if they appear in two consecutive positions in B_r.

We will show that any two neighbors are coprime (Claim 1) and any pair (i, j) of coprime positive integers that sum up to r appears exactly once as a pair of ordered neighbors in B_r (Claim 2). Together, these claims show that the neighbor pairs in B_{r-1} that sum up to r are exactly the pairs of coprime positive integers that sum up to r.

Fact 1. *If i and j are coprime then each of them is coprime with $i + j$ and with $i - j$.*

Proof. Suppose $\gcd(i, i + j) = d > 1$. Then d divides both i and $(i + j) - i = j$, implying that i and j are not coprime, which is a contradiction. Hence, i and $i + j$ are coprime. By symmetry, j and $i + j$ are coprime. Using a similar argument, we can show that each of i and j is coprime with $i - j$. □

By this fact, there is a bijection between pairs (i, j) of coprime positive integers that sum up to r and integers $i \in [r]$ that are coprime to r. Hence, the number of neighbor pairs in B_{r-1} that sum up to r is $\varphi(r)$. Therefore, B_r contains $\varphi(r)$ occurrences of r. By induction, it follows that $|B_r| = |B_{r-1}| + \varphi(r) = \Phi(r - 1) + 1 + \varphi(r) = \Phi(r) + 1$, proving the Lemma. □

We now prove the necessary claims.

Claim 1. *For all $r \geq 2$, if two numbers are neighbors in B_r, then they are coprime.*

Proof. We prove the claim by induction. For the base case of $r = 2$, the claim follows from the fact that 1 and 2 are coprime. For the general case of $r \geq 3$, recall that B_r was obtained from B_{r-1} by inserting an element r between all neighbors i and j in B_{r-1} that summed to r. By the induction hypothesis, $\gcd(i, j) = 1$, and, hence, by Fact 1, $\gcd(i, r) = \gcd(i, i + j) = 1$ and $\gcd(r, j) = \gcd(i + j, j) = 1$. Therefore, any two neighbors in B_r must be coprime. □

Claim 2. *For all $r \geq 3$, every ordered pair (i, j) of coprime positive integers that sum to r occurs exactly once as neighbors in B_{r-1}.*

Proof. We prove the claim by strong induction. The reader can verify the base case (when $r = 3$). For the inductive step, suppose the claim holds for all $k \leq r - 1$ for some $r \geq 4$. Consider an ordered pair (i, j) of coprime positive integers that sum to r. Assume that $i > j$; we know that $i \neq j$, and the case of $i < j$ is symmetric. Since $r \geq 4$, we have that $i \geq 3$. In the recursive construction of the sequences $\{B_k\}$, the elements i are added to the sequence B_i when B_i is created from B_{i-1}. Since $j < i$, all the elements j are already present in B_{i-1}. By Fact 1, since $\gcd(i, j) = 1$, we get that $\gcd(i - j, j) = 1$. By the inductive hypothesis, pair $(i - j, j)$ appears exactly once as an ordered pair of neighbors in B_{i-1}. Consequently, (i, j) must appear exactly once as an ordered pair of neighbors in B_i. No new elements i, j are added to the sequence in later stages, when $k > i$. Also, no new elements are inserted between i and j when $i + 1 \leq k \leq i + j - 1 = r - 1$. Therefore, the ordered neighbor pair (i, j) appears exactly once in B_{r-1}. □

3.2. Lower Bound

In this section, we prove Theorem 2.

Theorem 2. *For all $n \in \mathbb{N}$, the graph $C_{n,n}$ has readability $\Omega(\log n)$.*

First, we will need the notion of a HUB decomposition from [7]. Given $G = (V_s, V_p, E)$ and a function $w : E \to [k]$, we define G_i, for $i \in [k]$, as the graph with the same vertex set as G and edges given by $E(G_i) = \{e \in E \mid w(e) = i\}$. Observe that the edge sets of $G_1, \dots, G_k$ form a partition of E. We say that w is a *hierarchical-union-of-bicliques decomposition*, abbreviated as *HUB decomposition*, if the following conditions hold: i) for all $i \in [k]$, G_i is a disjoint union of bicliques, and ii) if two distinct vertices u and v are non-isolated twins in G_i for some $i \in \{2, \dots, k\}$ then, for all $j \in [i-1]$, u and v are (possibly isolated) twins in G_j. The parameter k is called the size of the HUB decomposition w. Now, consider an arbitrary HUB decomposition of $C_{n,n}$ and let h be its size. We will show that $h \geq \log n$.

Lemma 4. *For each $i \in \{0, \dots, h-1\}$, graph G_{h-i} has maximum degree at most 2^i.*

Proof. We prove the lemma by strong induction on i. The base case is when $i = 0$. Observe that if G_h has non-isolated twins, then those must be twins in G_j for each $j \in [h]$, and, as a result, in $C_{n,n}$. Since $C_{n,n}$ has no twins, G_h has no non-isolated twins. By the first property of the HUB decomposition, G_h must have maximum degree at most 1.

For general i, let F_i denote the graph $(V_s, V_p, \bigcup_{j\in\{0,\dots,i-1\}} E(G_{h-j}))$. By the inductive hypothesis, F_i has maximum degree at most $\sum_j 2^j = 2^i - 1$. Consider a group of vertices S in the same part of $C_{n,n}$ that have the same degree in the graph $C_{n,n} - E(F_i)$. Since no two vertices in the same part of $C_{n,n}$ have the same degree, no two vertices in S have the same degree in F_i. Combining this with the fact that the degree of any vertex in F_i is at most $2^i - 1$, we infer that $|S| \leq 2^i$.

By the second property of the HUB decomposition, if two vertices are non-isolated twins in G_{h-i}, they are twins in $C_{n,n} - E(F_i)$. Consequently, each group of twins in G_{h-i} has size at most 2^i. By the first property of the HUB decomposition, G_{h-i} is a disjoint union of bicliques. It follows that each of these bicliques is a subgraph of the complete bipartite graph $K_{2^i,2^i}$, thus implying the required bound on the maximum degree. ☐

Proof of Theorem 2. By Lemma 4, graph G_{h-i} has at most $2^i n$ edges. Since the edge sets of $G_1, \dots, G_h$ form a partition of the edge set of $C_{n,n}$, the number of edges in $C_{n,n}$ is $\frac{n(n+1)}{2} \leq \sum_{i=0}^{h-1} 2^i n = n(2^h - 1)$. We get that $h \geq \log_2(n+3) - 1$. It was shown in [7] that the readability of every bipartite graph G is bounded from below by the minimum size of a HUB decomposition of G. This completes the proof. ☐

4. A Characterization of Graphs with Readability at most 2

In this section, we characterize bipartite graphs with readability at most 2 by proving Theorem 3. Due to Lemma 1, it is enough to obtain such a characterization for connected twin-free bipartite graphs. We later use this characterization in Section 5 to develop a polynomial time algorithm for recognizing graphs of readability at most 2 and also in Section 6 to prove a lower bound on the readability of general grids. Recall that a domino is the graph obtained from C_6 by adding an edge between a pair of two vertices at distance 3 (see Figure 4). We first define the notion of a *conforming* matching, which is implicitly used in the statement of Theorem 3.

Definition 3. A matching M in a bipartite graph G is *conforming* if the following conditions are satisfied:

1. The graph $G' = G - M$ is a disjoint union of bicliques (equivalently: P_4-free).
2. For $U \subseteq V(G)$, if $G[U]$ is a C_6, then $G'[U]$ is the disjoint union of three edges.
3. For $U \subseteq V(G)$, if $G[U]$ is a domino, then $G'[U]$ is the disjoint union of a C_4 and an edge.

We prove Theorem 3 by showing that a bipartite graph G has readability at most 2 iff G has a conforming matching. We restate the theorem here for the convenience of the reader.

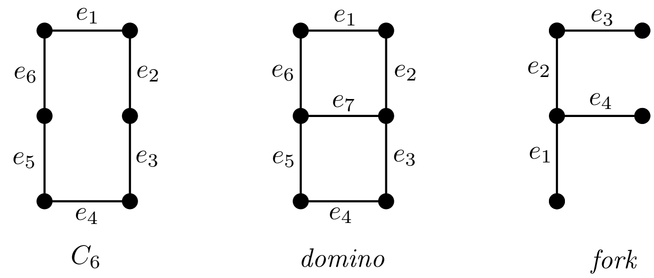


Figure 4: The C_6, the domino and the fork

Theorem 3. *A twin-free bipartite graph G has readability at most 2 if and only if G has a matching M such that the graph $G' = G - M$ satisfies the following properties:*

1. *G' is a disjoint union of complete bipartite graphs.*
2. *For $U \subseteq V(G)$, if $G[U]$ is a C_6, then $G'[U]$ is the disjoint union of three edges.*
3. *For $U \subseteq V(G)$, if $G[U]$ is a domino, then $G'[U]$ is the disjoint union of a C_4 and an edge.*

Proof. We show that $r(G) \leq 2$ if and only if G has a conforming matching.

Necessity. Suppose that $G = (V_s, V_p, E)$ is a twin-free bipartite graph of readability at most 2. Let ℓ be an overlap labeling of G of length at most 2. For vertices $u \in V_s, v \in V_p$, we denote by $\mathrm{ov}_\ell(u, v)$ the maximum length of a suffix of $\ell(u)$ that is also a prefix of $\ell(v)$. Since ℓ is an overlap labeling of G, we can partition the edge set of G into two sets, E_1 and E_2, by setting $E_1 = \{(u, v) \in E \mid \mathrm{ov}_\ell(u, v) = 1\}$ and $E_2 = E \setminus E_1$. Then for all $(u, v) \in E_2$, we have $\mathrm{ov}_\ell(u, v) = 2$, that is, $\ell(u) = \ell(v)$. Note that due to the definition of the overlap function, for every edge $(u, v) \in E_2$, the labels of u and v must not have an overlap of length one.

We claim that E_2 is a conforming matching. If E_2 is not a matching, we can assume by symmetry that there exists a vertex $u \in V_s$ and a pair of distinct vertices v, w in V_p such that $\{(u, v), (u, w)\} \subseteq E_2$. But then $\ell(v) = \ell(u) = \ell(w)$, which implies that v and w are twins in G, a contradiction. Thus, E_2 is a matching.

Let G' denote the graph $G - E_2$. Next, we show that G' is P_4-free. If (u, v, x, y) forms an induced P_4 in (V, E_1) (with edge set $\{(u, v), (x, v), (x, y)\}$), then $\mathrm{suf}_1(\ell(u)) = \mathrm{pre}_1(\ell(v)) = \mathrm{suf}_1(\ell(x)) = \mathrm{pre}_1(\ell(y))$, implying that $(u, y) \in E_1$, a contradiction. Therefore, G' is P_4-free.

Now let us verify the remaining two properties in the definition of a conforming matching. Let U be a subset of vertices in G. If $G[U]$ is isomorphic to C_6, we would like to show that $G'[U]$ is a union of three disjoint edges. Suppose for the sake of contradiction that it is not. Consider an edge labeling of $G[U]$ as in Figure 4. Since E_2 is a matching, the only other way for G' to be P_4-free, i.e., if it was not a union of three disjoint edges, is for E_2 to contain two diametrically opposite edges of $G[U]$, say e_1 and e_4. Let $e_i = (x_i, x_{i+1})$ for all $i \in [6]$ (addition modulo 6). Let, without loss of generality, $x_1 \in V_s$.Then $x_2 \in V_p$. Since $e_1 \in E_2$ by our assumption, we have $\ell(x_1) = \ell(x_2)$, say $\ell(x_1) = \ell(x_2) = ab$. We have $\mathrm{suf}_1(\ell(x_5)) = \mathrm{pre}_1(\ell(x_6)) = \mathrm{suf}_1(\ell(x_1)) = b$ and $\mathrm{pre}_1(\ell(x_4)) = \mathrm{suf}_1(\ell(x_3)) = \mathrm{pre}_1(\ell(x_2)) = a$. Since $e_4 \in E_2$, we get $\ell(x_4) = \ell(x_5) = ab$. Therefore $\ell(x_1) = \ell(x_4)$, which is a contradiction, since $(x_1, x_4) \notin E$ and ℓ is an overlap labeling of G.

Finally, suppose that $G[U]$ is isomorphic to the domino, and assume an edge labeling as in Figure 4. Since G' is P_4-free, $G'[U]$ is also P_4-free and hence $G'[U]$ can only be isomorphic to either (1) a disjoint union of a C_4 and an edge (which is what we want to show), or (2) a disjoint union of two P_3's. Suppose we are in case (2). Then we have $e_1, e_4, e_7 \in E_2$. Let $e_i = (x_i, x_{i+1})$ for all $i \in \{1, \ldots, 6\}$ (addition modulo 6). We may assume without loss of generality that $x_1 \in V_s$. Since $e_1, e_4 \in E_2$ and $e_2, e_3, e_5, e_6 \in E_1$, we can follow the same reasoning as above, and conclude that the labels of x_1 and x_4 are equal, which is a contradiction, since $(x_1, x_4) \notin E$ and ℓ is an overlap labeling of G. This establishes the necessity of the condition.

Sufficiency. Suppose now that $G = (V_s, V_p, E)$ is a twin-free bipartite graph with a conforming matching M. We will show that G has readability at most 2 by constructing an overlap labeling of G of length at most 2. Since M is a conforming matching, the graph $G' = G - M$ is P_4-free, that is, a disjoint union of bicliques. Let $\{A_1, B_1\}, \ldots, \{A_k, B_k\}$ be the bipartitions of the vertex sets of the connected components (bicliques) $G_1, \ldots, G_k$ of G' (so that $A_i = V(G_i) \cap V_s$ for all i; some of the A_i's or B_i's may be empty). Then $\cup_{i=1}^{k} V(G_i) = V$. Assign a partial labeling over $\Sigma = \{1, \ldots, k\}$ to vertices of G by setting $\ell(v) = i$ if and only if $v \in V(G_i)$. For each edge $(u, v) \in M$, extend the labels of $u \in V_s$ and $v \in V_p$ as follows. Let $u \in A_i$ and $v \in B_j$. Then $i \neq j$ because edges of bicliques in $G - M$ cannot be in M. Replace $\ell(u) = i$ with $\ell(u) = ji$, and $\ell(v) = j$ with $\ell(v) = ji$. Since M is a matching, every vertex will have a label of length 1 or 2 at the end of this procedure. Extend the labels of length 1 by unique new characters to make them of length 2. By construction, the overlaps of the obtained labeling create all edges of $E(G') \cup M = E(G)$.

Let us verify that no new edges were created by ℓ. Suppose that u, v is a pair of vertices with with $u \in V_s$ and $v \in V_p$ and $\mathrm{ov}_\ell(u, v) > 0$. If $\ell(u)$ and $\ell(v)$ have an overlap of length 1, then $(u, v) \in E(G')$ by construction. Suppose that $\ell(u)$ and $\ell(v)$ do not have an overlap of length 1 but have an overlap of length 2. Then $\ell(u) = \ell(v) = ij$ for two distinct $i, j \in \Sigma$. By construction, vertex u is adjacent to a unique vertex w via a matching edge in M, moreover $u \in A_j$ and $w \in B_i$. If $w = v$, then the edge (u, v) is in M and hence in G. So we may assume that $w \neq v$. Similarly, vertex v is adjacent to a unique vertex z in M, and $z \in A_j$ and $v \in B_i$. If $u = z$, then again the edge (u, v) is in M and hence in G. So we may assume that $u \neq z$. Since $|A_j| \geq 2$, there exists a vertex $s \in B_j$. Similarly, since $|B_i| \geq 2$, there exists a vertex $t \in A_i$. Notice that $(u, v) \notin M$ since u is of degree 1 in M, and $(u, v) \notin E(G')$ since u and v belong to distinct connected components of G'. Therefore, $(u, v) \notin E(G)$, and, similarly, $(z, w) \notin E(G)$. But now, the subset $\{s, t, u, v, w, z\}$ induces a subgraph of G isomorphic to either a C_6 (if $(t, s) \notin E(G)$) or a domino (otherwise). In either case, one of the conditions for the C_6 and for the domino in Definition 3 is violated, contrary to the fact that M is a conforming matching.

This shows that ℓ is an overlap labeling of G and implies that the readability of G is at most 2. □

Corollary 2. *Every bipartite graph G of maximum degree at most 2 has readability at most 2.*

Proof. If G is a connected twin-free bipartite graph of maximum degree at most 2, then G is a path or an (even) cycle. In this case, the edge set of G can be decomposed into two matchings M_1 and M_2 by picking alternate edges. Both M_1 and M_2 are conforming matchings. Thus, by Theorem 3, G has readability at most 2. □

5. An Efficient Algorithm for Readability 2

In this section, we prove Theorem 4 by developing a polynomial time algorithm for the following problem.

Readability 2	
Instance:	A bipartite graph $G = (V_s, V_p, E)$.
Question:	Is $r(G) \leq 2$?

First, we use Lemma 1 and Corollary 2 to reduce the problem to connected twin-free bipartite graphs of maximum degree at least 3. We then apply Theorem 3 and reduce the problem to checking for the existence of a conforming matching (Definition 3). Finally, we show how to reduce this problem to the 2SAT problem (Lemma 5), which is well known to be solvable in linear time (see, e.g., [1]).

Theorem 4. *There exists an algorithm that, given a bipartite graph G, decides in polynomial time whether G has readability at most 2.*

Proof. Given a bipartite graph G, we first reduce the problem to its connected components. That is, if G is not connected, then, by Lemma 1(b), $r(G) \leq 2$ if and only if all components G' of G satisfy $r(G') \leq 2$. Second, assuming G is connected, we compute the twin-free reduction G' of G, which, by Lemma 1(d), does not change the readability. We test whether G' is of maximum degree at most 2. If this is the case, then, by Corollary 2, we assert that G has readability at most 2.

Consider a connected twin-free bipartite graph $G = (V, E)$ of maximum degree at least 3. Let E' denote the set of all edges $e = (u, v)$ in G such that either (1) $\{u, v\} \cup N(u) \cup N(v)$ has a vertex of degree at least

3, or (2) e is contained in some induced C_6. The definition of E' and the fact that G is connected and of maximum degree at least 3 imply that if an induced subgraph H of G is isomorphic to a C_4, a fork, a C_6, or a domino (see Figure 4), then $E(H) \subseteq E'$.

Let $X = \{x_e \mid e \in E'\}$ be a set of variables. We now define a 2SAT formula φ over X such that G has a conforming matching (and hence, readability at most 2) if and only if φ is satisfiable. The formula φ contains the following five types of clauses.

1. For each pair $\{e, f\} \subseteq E'$ of distinct edges that share an endpoint, add the clause $\overline{x_e} \vee \overline{x_f}$ to φ.
2. For each induced subgraph H of G isomorphic to C_4 and each matching $\{e, f\}$ in H, add the clauses $\overline{x_e} \vee x_f$ and $\overline{x_f} \vee x_e$ (equivalent to $x_e \leftrightarrow x_f$) to φ.
3. For each induced subgraph H of G isomorphic to C_6, with edges labeled as in Figure 4, add the clause $x_{e_1} \vee x_{e_2}$, the clauses corresponding to $x_{e_1} \leftrightarrow x_{e_3}$ and $x_{e_3} \leftrightarrow x_{e_5}$, and the clauses corresponding to $x_{e_2} \leftrightarrow x_{e_4}$ and $x_{e_4} \leftrightarrow x_{e_6}$ to φ.
4. For each induced subgraph H of G isomorphic to the domino, with edges labeled as in Figure 4, add the clauses $x_{e_2} \vee x_{e_3}$ and $x_{e_5} \vee x_{e_6}$ to φ.
5. For each induced subgraph H of G isomorphic to the fork, with edges labeled as in Figure 4, add the clause $x_{e_2} \vee x_{e_3}$ to φ.

The following lemma shows that if φ is satisfiable, then $r(G) \leq 2$, otherwise, $r(G) > 2$.

Lemma 5. *Graph G has a conforming matching if and only if formula φ is satisfiable.*

Proof. Suppose first that G has a conforming matching, say M. Let a be an assignment of Boolean values to the variables in X such that for every $e \in E'$, variable x_e is true if and only if $e \in M$. We will prove that a is a satisfying assignment for φ. It is easy to see that clauses of type (1) in φ are satisfied as M is a matching.

Consider a pair of clauses $\overline{x_e} \vee x_f$ and $\overline{x_f} \vee x_e$ of type (2) in φ. These correspond to an induced subgraph H of G isomorphic to a C_4 and a matching $\{e, f\}$ in H. Since M is a conforming matching, the graph $G - M$ is P_4-free, and so we have $e \in M$ if and only if $f \in M$. Hence a satisfies both the clauses.

Clauses in φ of type (3) deal with induced 6-cycles and those of type (4) deal with induced dominos. Both types of clauses are satisfied by a due to the fact that M, which is a conforming matching, satisfies conditions 2 and 3 in Definition 3.

Finally, clauses in φ of type (5) are satisfied only if for each induced subgraph H of G isomorphic to the fork (with edges labeled as in Figure 4), we have $\{e_2, e_3\} \cap M \neq \emptyset$. Suppose for the sake of contradiction that there exists an induced fork H for which this is not the case. Since $G - M$ is P_4-free, so is $H - M$ and hence e_1 and e_4 are both in M, which is a contradiction. This shows that formula φ is satisfiable.

For the converse direction, suppose that formula φ is satisfiable and let a be a satisfying assignment. Let M' be the set of edges $e \in E'$ such that x_e is set to true in a. Extend M' greedily to a set of edges M by setting $M = M'$ and then iteratively adding the middle edge of any induced subgraph H of G isomorphic to P_4 that contains no edge of M. We claim that the so obtained set M is a conforming matching of G. This will be easy to show once we prove the following claim.

Claim 3. *M is a matching in G with $M \cap E' = M' \cap E'$.*

Proof. The claim is true if $M = M'$, since M' is a matching by virtue of type (1) clauses. Henceforth, assume that $M \neq M'$. We will first show that $M \cap E' = M' \cap E'$. For this, it is enough to prove that $e \notin E'$ for each $e \in M \setminus M'$.

Consider an edge $e \in M \setminus M'$. By our construction of M, the edge e is the middle edge of an induced subgraph H of G isomorphic to P_4 such that H contains no other edge of M. In particular, H has no edge of M'. Let u and v be the endpoints of e, and let x and y be the remaining two vertices in H such that (x, u) and (v, y) are the other two edges in H. Assume for the sake of contradiction that $e \in E'$. Then, either (a) $\{u, v\} \cup N(u) \cup N(v)$ has a vertex of degree at least 3, or (b) e is contained in some induced C_6.

Suppose first that (b) holds. Then, by virtue of the type (3) clauses, either $(u, v) \in M'$ or both (x, u) and (y, v) are in M'. Both cases contradict our premise that H contains no edge of M'.

Suppose now that (a) holds. Assume that the degree of u is at least 3. Let w be a neighbor of u such that $w \neq x$. We will show that the set $\{x, u, w, v, y\}$ induces a fork in G. Since G is a bipartite graph, it has no C_3's and hence $(w, x), (w, v) \notin E$. If $(w, y) \in E$, then the set $\{w, u, v, y\}$ induces a C_4. Since u is of degree at least 3, we have $\{(w, u), (u, v), (v, y), (y, w)\} \subseteq E'$ and hence, by virtue of clauses of type (2), either (u, v) and (w, y) are in M', or both (u, w) and (v, y) are in M'. Both of these contradict our premise that H contains no edge of M'. Therefore, the set $\{x, u, w, v, y\}$ induces a fork in G, and by virtue of its associated type (5) clause, either (u, v) or (v, y) is in M'. This contradicts our assumption that H does not have any edge in M'. Thus, the degree of u is 2. By a symmetric argument, the degree of v is 2. Thus, the only way for (a) to hold is for either $x \in N(u)$ or $y \in N(v)$ to have degree at least 3. By symmetry, we may assume that x has degree at least 3. Let $s, t \in N(x) \setminus \{u\}$. Since v is of degree 2, it is non-adjacent to both s and t, hence the set $\{s, t, x, u, v\}$ induces a fork in G and hence either (x, u) or (u, v) is in M', a contradiction. Thus, we have proved that if $e \in M \setminus M'$, then $e \notin E'$ and therefore, $M \cap E' = M' \cap E'$.

We will now show that M is a matching. From the above arguments, we know that for each edge $(u, v) = e \in M \setminus M'$, degree of both u and v are at most 2. Thus, the only edges adjacent with e are the ones that form the induced copy of P_4 with it. As neither of them are in M, the edge e does not share an endpoint with any other edge of M. This completes the proof of the claim. □

It remains to verify that M is a conforming matching. First, suppose for the sake of contradiction that $G - M$ is not P_4-free. Fix an induced P_4 in $G - M$, say H, with edges $\{(u, v), (v, w), (w, x)\}$. The set $V(H) = \{u, v, w, x\}$ does not induce a P_4 in G, for otherwise, we would have added one of the edges of H to M. Thus, we have, $(u, x) \in E$, implying that, $(u, x) \in M$. Recall that since G is connected and of maximum degree at least 3, the set $V(H)$ contains a vertex of degree at least 3 in G, which implies that all edges of the C_4 induced by $V(H)$ must be in E'. Since $M \cap E' = M' \cap E'$, we have in particular $(u, x) \in M'$. Since (u, x) is the only M'-edge in the C_4 induced by $V(H)$ in G, it contradicts the fact that the type (2) clause corresponding to that C_4 is satisfied by the assignment a.

Second, let H be an induced subgraph of G isomorphic to C_6. By the definition of E', we have that $E(H) \subseteq E'$ and consequently $M \cap E(H) = M' \cap E(H)$. The fact that the clauses of type (3) corresponding to H are satisfied by a implies that $H - M$ is a union of three disjoint edges.

Finally, let H be an induced subgraph of G isomorphic to the domino (with edges labeled as in the right side of Figure 4). By the definition of E', we again have $E(H) \subseteq E'$ and thus $M \cap E(H) = M' \cap E(H)$. The fact that the clauses of type (4) corresponding to H are satisfied by a implies that $M' \cap \{e_2, e_3\} \neq \emptyset$ and $M' \cap \{e_5, e_6\} \neq \emptyset$. We may assume by symmetry that $M' \cap \{e_2, e_3\} = \{e_2\}$. The fact that the clause of type (2) is satisfied corresponding to the C_4 with edge set $\{e_1, e_2, e_6, e_7\}$ and the 2-matching $\{e_2, e_6\}$ implies that $e_6 \in M'$. Consequently, $M \cap E(H) = M' \cap E(H) = \{e_2, e_6\}$ and the desired condition holds. This proves that M is a conforming matching in G and completes the proof of the lemma. □

The correctness of the algorithm follows from Theorem 3. We can compute formula φ from a given graph G in polynomial time. The 2SAT problem is solvable in linear time [1], and clearly, all the other steps of the algorithm can be implemented to run in polynomial time. The method given above can easily be modified so that it also efficiently computes an overlap labeling of length at most 2 in case of a yes instance. □

6. Readability of Grids and Grid Graphs

In this section, we determine the readability of grids by proving Theorem 5. We first look at toroidal grids, which are closely related to grids. For positive integers $m \geq 3$ and $n \geq 3$, the *toroidal grid* $TG_{m,n}$ is obtained from the grid $G_{m,n}$ by adding edges $((i, 0), (i, n-1))$ and $((0, j), (m-1, j))$ for all $i \in \{0, \ldots, m-1\}$ and $j \in \{0, \ldots, n-1\}$. (See Figure 2 for an example.) The graph $TG_{m,n}$ is bipartite if and only if m and n are both even. In this case, a bipartition can be obtained by setting $V(TG_{m,n}) = V_s \cup V_p$ where $V_s = \{(i, j) \in V(TG_{m,n}) : i + j \equiv 0 \pmod 2\}$ and $V_p = \{(i, j) \in V(TG_{m,n}) : i + j \equiv 1 \pmod 2\}$.

Lemma 6. *For all integers $n > 0$, we have $r(TG_{4n,4n}) \leq 3$.*

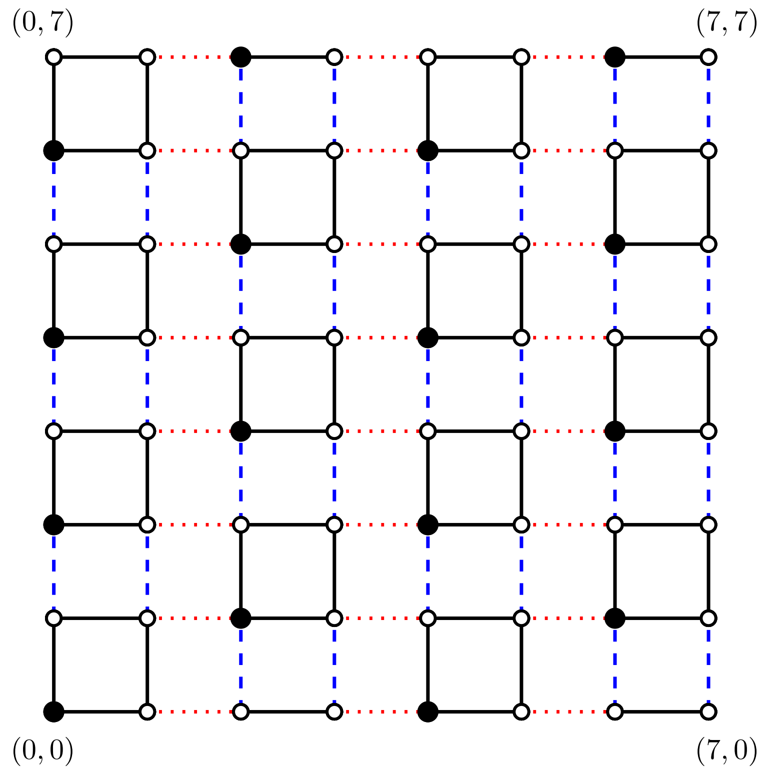


Figure 5: The graph $TG_{8,8}$ (with edges between the extreme layers omitted in the drawing) where the lower left endpoints of the squares in G_0 are marked using solid dots, and the edges in M_1 and M_2 are drawn using (red) dotted and (blue) dashed lines, respectively.

Proof. Fix n and let $G = TG_{4n,4n}$. Each vertex u of G has associated coordinates (u_1, u_2), where $u_1, u_2 \in \{0, 1, \dots, 4n-1\}$. All arithmetic on coordinates will be performed modulo $4n$. By Lemma 1(c), we may assume without loss of generality the bipartition (V_s, V_p) given above.

We decompose G into three subgraphs. The first subgraph consists of squares. A *square* S_u for a vertex u of G is the subgraph of G induced by vertices $\{u, u+(0,1), u+(1,0), u+(1,1)\}$. The subgraph G_0 of G is the union of all squares S_u, where either (1) u_1 is divisible by 4 and u_2 is divisible by 2, or (2) u_1+2 is divisible by 4 and u_2+1 is divisible by 2. Note that G_0 has $4n^2$ squares.

We assign each square a unique identifier from the range $\{0, 1, \dots, 4n^2-1\}$. Observe that each vertex u of G belongs to exactly one square in G_0, and we use $\ell_0(u)$ to denote the identifier of the square in G_0 to which u belongs. We divide the edges of G into *horizontal* and *vertical* ones respectively, according to whether they connect a pair of vertices that differ in their first, resp., second coordinates. Next, we define M_1 (respectively, M_2) to be the set of all horizontal (respectively, vertical) edges of $G - E(G_0)$. For $i \in \{1, 2\}$, we use $M_i(u)$ to denote the vertex matched to u in M_i. Figure 5 illustrates the graph $TG_{8,8}$ (without the wrap-around edges, for simplicity) where the lower left endpoints of the squares in G_0 are marked using black dots, and the edges in M_1 and M_2 are drawn using dotted and dashed lines, respectively.

We now define a labeling ℓ of G. For each vertex u of G, define $\ell_1(u) = \ell_0(M_1(u))$ and $\ell_2(u) = \ell_0(M_2(u))$. If $u \in V_s$ then define $\ell(u) = \ell_2(u)\ell_1(u)\ell_0(u)$; if $u \in V_p$ then define $\ell(u) = \ell_0(u)\ell_1(u)\ell_2(u)$, i.e., the same characters, but in reverse order. The following claim shows that ℓ is an overlap labeling and, since ℓ has labels of length 3, proves the lemma.

Claim 4. *Labeling ℓ is an overlap labeling of G.*

Proof. First, we make three observations about G_0, M_1, and M_2.

Observation 1. Each edge of G is in exactly one of $E(G_0), M_1$, and M_2. Both M_1 and M_2 are perfect matchings in G.

Observation 2. If $(u,v) \in M_2$ then $\ell_0(M_1(u)) = \ell_0(M_1(v))$, i.e., M_1 matches u and v to vertices in the same square of G_0.

Observation 3. Any pair of squares in G_0 is connected by at most one edge in M_1. That is, for all pairs (id_1, id_2) of square ids, at most one edge $(u,v) \in M_1$ satisfies $\ell_0(u) = id_1$ and $\ell_0(v) = id_2$.

First, we show that, for every edge (u,v) of G, where $u \in V_s$ and $v \in V_p$, the label $\ell(u)$ overlaps the label $\ell(v)$. By Observation 1, each edge of G is in one of G_0, M_1 and M_2. If (u,v) is in G_0, then u and v belong to the same square of G and, by construction $\ell_0(u) = \ell_0(v)$. That is,

$$\mathrm{suf}_1(\ell(u)) = \ell_0(u) = \ell_0(v) = pre_1(\ell(v)).$$

If $(u,v) \in M_1$, then $\ell_1(u) = \ell_0(v)$ and $\ell_1(v) = \ell_0(u)$, by the definition of ℓ_1. Therefore,

$$\mathrm{suf}_2(\ell(u)) = \ell_1(u)\ell_0(u) = \ell_0(v)\ell_1(v) = \mathrm{pre}_2(\ell(v)).$$

If $(u,v) \in M_2$, then $\ell_2(u) = \ell_0(v)$ and $\ell_2(v) = \ell_0(u)$, by the definition of ℓ_2. By Observation 2 and the definition of ℓ_1, we get that $\ell_1(u) = \ell_1(v)$. That is,

$$\ell(u) = \ell_2(u)\ell_1(u)\ell_0(u) = \ell_0(v)\ell_1(v)\ell_2(v) = \ell(v).$$

In all three cases $\ell(u)$ overlaps $\ell(v)$.

It remains to show that if, for $u \in V_s$ and $v \in V_p$, label $\ell(u)$ overlaps label $\ell(v)$ then (u,v) is an edge in G. Since labels $\ell(u)$ and $\ell(v)$ have length 3, the overlap from $\ell(u)$ to $\ell(v)$ can be of length 1, 2 or 3. If $\mathrm{suf}_1(\ell(u)) = \mathrm{pre}_1(\ell(v))$ then $\ell_0(u) = \ell_0(v)$, that is, u and v are in the same square of G_0. Hence, (u,v) is an edge in G_0 and, consequently, in G.

If $\mathrm{suf}_2(\ell(u)) = \mathrm{pre}_2(\ell(v))$ then $\ell_1(u) = \ell_0(v)$ and $\ell_0(u) = \ell_1(v)$. By the definition of ℓ_1, this implies that both $(u, M_1(u))$ and $(M_1(v), v)$ connect squares of G_0 with identifiers $\ell_0(u)$ and $\ell_0(v)$. By Observation 3, $(u, M_1(u))$ is the same edge as $(M_1(v), v)$, namely, (u,v). Hence, (u,v) is in M_1 and, consequently, in G.

Finally, suppose $\ell(u) = \ell(v)$. Then $\ell_2(u)\ell_1(u)\ell_0(u) = \ell_0(v)\ell_1(v)\ell_2(v)$. Since $\ell_2(u) = \ell_0(v)$ and $\ell_0(u) = \ell_2(v)$, it follows that $(u, M_2(u))$ and $(M_2(v), v)$ are vertical edges connecting the same pair of squares in G_0. Since $\ell_1(u) = \ell_1(v)$, we have that $M_1(u)$ and $M_1(v)$ belong to the same square in G_0. Both conditions can hold only if $(u, M_2(u))$ and $(M_2(v), v)$ are the same edge, namely, (u,v). Hence, (u,v) is in M_2 and, consequently, in G. In all cases, we proved that (u,v) is an edge of G. □

This completes the proof of Lemma 6. □

We can now prove Theorem 5, determining the readability of $G_{m,n}$. We first recall the following simple observation (which follows, e.g., from [7, Theorem 4.3]).

Lemma 7. *A bipartite graph G has: (i) $r(G) = 0$ if and only if G is edgeless, and (ii) $r(G) \leq 1$ if and only if G is P_4-free (equivalently: a disjoint union of bicliques).*

Theorem 5. *For any two positive integers m, n with $m \leq n$, we have*

$$r(G_{m,n}) = \begin{cases} 3, & \text{if } m \geq 3; \\ 2, & \text{if } (m = 2 \text{ and } n \geq 3) \text{ or } (m = 1 \text{ and } n \geq 4); \\ 1, & \text{if } (m,n) \in \{(1,2),(1,3),(2,2)\}; \\ 0, & \text{if } m = n = 1. \end{cases}$$

Proof. First, by Lemma 7, $r(G_{m,n})$ is 0 if $m = n = 1$ and positive, otherwise. Second, when $(m,n) \in \{(1,2),(1,3),(2,2)\}$, the graphs $G_{m,n}$ are isomorphic to $K_{1,1}, K_{1,2}$, and $K_{2,2}$, respectively. Thus, by Lemma 7, their readability is 1.

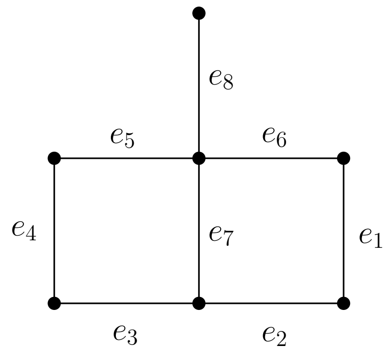


Figure 6: The graph F.

Third, when $m + n \geq 5$, the grid $G_{m,n}$ contains an induced P_4, implying that $r(G_{m,n}) \geq 2$. By Theorem 3, a twin-free bipartite graph G has readability at most 2 if and only if G has a conforming matching. (See Definition 3.) When $m + n \geq 5$, the grid $G_{m,n}$ is twin-free. If $m = 2$ and $n \geq 3$, then $M = \{((i,j),(i,j+1)) \mid i \in \{0,1\} \text{ and } j \in \{0,\ldots,n-2\} \text{ is even}\}$ is a conforming matching in $G_{m,n}$, so $r(G_{m,n}) = 2$. If $m = 1$ and $n \geq 4$, then $G_{m,n}$ is isomorphic to a path of length at least three. Since its maximum degree is 2, we get $r(G_{m,n}) \leq 2$, by Corollary 2. Thus, $r(G_{m,n}) = 2$.

To show that $r(G_{m,n}) \leq 3$ for $m \geq 3$ and $n \geq 3$, we observe that $G_{m,n}$ (for $m \leq n$) is an induced subgraph of $TG_{4n,4n}$. By Lemmas 1(a) and 6, we have that $r(G_{m,n}) \leq r(TG_{4n,4n}) \leq 3$.

To show that $r(G_{m,n}) \geq 3$, let F be the graph obtained by taking the graph $G_{3,2}$ and adding a new vertex adjacent to one of the degree-3 vertices of $G_{3,2}$; see Figure 6.

Clearly, F is a bipartite graph and an induced subgraph of $G_{m,n}$. Since F is also twin-free, we can prove that $r(F) > 2$ by applying Theorem 3, provided we show that F does not have a conforming matching. Assume the edge labeling as in Figure 5(a) and suppose for a contradiction that F has a conforming matching M. The third condition in Definition 3 implies that $M \cap (E(F) \setminus \{e_8\}) \in \{\{e_2, e_6\}, \{e_3, e_5\}\}$. By symmetry, we may assume that $M \cap (E(F) \setminus \{e_8\}) = \{e_2, e_6\}$. Since M is a matching, we have $e_8 \notin M$. But now the graph $F - M$ contains an induced P_4 with edge set $\{e_4, e_5, e_8\}$, a contradiction to the fact that M is conforming. This shows that $r(F) \geq 3$. By Lemma 1(a), $r(G_{m,n}) \geq r(F) \geq 3$ if $m \geq 3$ and $n \geq 3$. □

7. Conclusion

In this work we gave several results on families of n-vertex bipartite graphs with readability $o(n)$. The results were obtained by developing new or applying a variety of known techniques to the study of readability. These include a graph-theoretic characterization in terms of matchings, a reduction to 2SAT, an explicit construction of overlap labelings analyzed via number theoretic notions, and a new lower bound applicable to dense graphs with a large number of distinct degrees. One of the main specific questions left open by our work is to close the gap between the $\Omega(\log n)$ lower bound and the $\mathcal{O}(\sqrt{n})$ upper bound on the readability of n-vertex bipartite chain graphs. For general graphs, it remains open whether the problem of computing the readability of a given bipartite graph is NP-hard, and whether the decision version of the problem is in NP. Other questions related to readability include determining the computational complexity of recognizing bipartite graphs of readability at most k, where k is a constant greater than 2, studying the parameter from an approximation point of view, and relating it to other graph invariants. For instance, for a positive integer k, what is the maximum possible readability of a bipartite graph of maximum degree at most k? Another interesting direction would be to study the complexity of various computational problems on graphs of low readability.

Acknowledgments. The authors are grateful to two anonymous reviewers for their helpful remarks. The result of Section 3.1 was discovered with the help of The On-Line Encyclopedia of Integer Sequences ® [23]. This work has been supported in part by NSF awards DBI-1356529, CCF-1439057, IIS-1453527, and IIS-1421908 to P.M. and by the Slovenian Research Agency (I0-0035, research program P1-0285 and research projects N1-0032, N1-0102, J1-6720, and J1-7051) to M.M. The authors S.R. and N.V. were supported in part by NSF grants CCF-1422975 and CCF-1832228 to S.R. The author N.V. was also supported by Pennsylvania State University College of Engineering Fellowship and Pennsylvania State University Graduate Fellowship and in part by NSF grant IIS-1453527 to P.M. V.J. did most of his work on the paper while he was an undergraduate student at the University of Primorska. The main idea of the proof of Lemma 6 was developed in his final project paper [14].

References

[1] Bengt Aspvall, Michael F. Plass, and Robert Endre Tarjan. A linear-time algorithm for testing the truth of certain quantified Boolean formulas. *Inform. Process. Lett.*, 8(3):121–123, 1979.

[2] Jacek Błażewicz, 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.

[3] Jacek Błażewicz, 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.

[4] Jacek Blazewicz, Alain Hertz, Daniel Kobler, and Dominique de Werra. On some properties of dna graphs. *Discrete Applied Mathematics*, 98(1):1–19, 1999.

[5] 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.

[6] Rayan Chikhi, Vladan Jovicic, Stefan Kratsch, Paul Medvedev, Martin Milanic, Sofya Raskhodnikova, and Nithin M. Varma. Bipartite graphs of small readability. In *Computing and Combinatorics - 24th International Conference, COCOON 2018, Qing Dao, China, July 2-4, 2018, Proceedings*, pages 467–479, 2018.

[7] Rayan Chikhi, Paul Medvedev, Martin Milanič, and Sofya Raskhodnikova. On the readability of overlap digraphs. *Discrete Applied Mathematics*, 205:35–44, 2016.

[8] Brent N. Clark, Charles J. Colbourn, and David S. Johnson. Unit disk graphs. *Discrete Mathematics*, 86(1-3):165–177, 1990.

[9] Josep Díaz, Mathew D. Penrose, Jordi Petit, and Maria J. Serna. Approximating layout problems on random geometric graphs. *J. Algorithms*, 39(1):78–116, 2001.

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

[11] Godfrey H. Hardy and Edward M. Wright. *An Introduction to the Theory of Numbers*. The Clarendon Press, Oxford University Press, New York, fifth edition, 1979.

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

[13] Alon Itai, Christos H. Papadimitriou, and Jayme Luiz Szwarcfiter. Hamilton paths in grid graphs. *SIAM J. Comput.*, 11(4):676–686, 1982.

[14] Vladan Jovičić. *Odčitljivost digrafov in dvodelnih grafov (Readability of digraphs and bipartite graphs)*. 2016. Final project paper (in English). University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies, Koper, Slovenia, 2016. Available online as arXiv:1612.07113 [cs.DM] at `https://arxiv.org/abs/1612.07113`.

[15] Marta Kasprzak. Classification of de Bruijn-based labeled digraphs. *Discrete Applied Mathematics*, 2016.

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

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

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

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

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

[21] Rudi Pendavingh, Petra Schuurman, and Gerhard J Woeginger. Recognizing dna graphs is difficult. *Discrete applied mathematics*, 127(1):85–94, 2003.

[22] Jared T Simpson and Richard Durbin. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, 2011.

[23] Neil J. A. Sloane. The On-Line Encyclopedia of Integer Sequences, 2016. Published electronically at `https://oeis.org`.

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

[25] Douglas B. West. *Introduction to Graph Theory*. Prentice Hall, Inc., Upper Saddle River, NJ, 1996.