Bipartite Graphs of Small Readability*

Rayan Chikhi[1], Vladan Jovičić[2], Stefan Kratsch[3], Paul Medvedev[4], Martin Milanič[5], Sofya Raskhodnikova[6], and Nithin Varma[6]

[1] CNRS, UMR 9189, rayan.chikhi@univ-lille1.fr
[2] ENS Lyon, France, vladan94.jovicic@gmail.com
[3] Institut für Informatik, Humboldt-Universität zu Berlin
kratsch@informatik.hu-berlin.de
[4] The Pennsylvania State University, USA pashadag@cse.psu.edu
[5] IAM and FAMNIT, University of Primorska, Koper, Slovenia
martin.milanic@upr.si
[6] Boston University, USA {sofya, nvarma}@bu.edu

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 (Braga and Meidanis, *LATIN* 2002) is exponential in the maximum degree of the graph. Graphs that arise in bioinformatic 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 proof of the upper bound on 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.

1 Introduction

In this work we further the study of *readability* of bipartite graphs initiated by Chikhi et al. [6]. 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

* The full version of this paper is available online [5].

nonnegative integer r such that there is an overlap labeling of G of length r. 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 [4] implies that these variants are within constant factors of each other, where the constants are logarithmic in the alphabet sizes.

The notion of readability arises in the study of overlap digraphs. Overlap digraphs constructed from DNA strings have various applications in bioinformatics.[7] Most of the graphs that occur as the overlap graphs of genomes have low readability. Chikhi et al. [6] show that the readability of overlap digraphs is asymptotically equivalent to that of balanced bipartite graphs: there is a bijection between overlap digraphs and balanced bipartite graphs that preserves readability up to (roughly) a factor of 2. This motivates 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 in a paper on overlap digraphs by Braga and Meidanis [4]. As observed by Chikhi et al. [6], it follows from [4] 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. Chikhi et al. [6] showed that almost all bipartite graphs with n vertices in each part have readability $\Omega(n/\log n)$. They also constructed an explicit graph family (called Hadamard graphs) with readability $\Omega(n)$.

For trees, readability can be defined in terms of an extremal question on certain integer functions on the edges, without any reference to strings or their overlaps [6]. 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. [6] developed a lower bound technique for graphs where the overlap between the neighborhoods of any two vertices is limited. 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 [6]), whereas the problem of recognizing graphs of readability k is open for all $k \geq 3$.

We also give a formula for the readability of grids, showing in particular that it never exceeds 3. As a corollary, we obtain a polynomial-time algorithm to determine the readability of induced subgraphs of grids.

[7] In the context of genome assembly, variants of overlap digraphs appear as either de Bruijn graphs [11] or string graphs [18, 21] and are the foundation of most modern assemblers (see [17, 19] for a survey). Several graph-theoretic parameters of overlap digraphs have been studied [2, 1, 3, 9, 15, 16, 20, 23], with a nice survey in [14].

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. All missing proofs can be found in the full version [5].

To study readability, it suffices to consider bipartite graphs that are connected and *twin-free*, i.e., no two nodes in the same part have the same sets of neighbors [6]. As connected bipartite graphs have a unique bipartition up to swapping the two parts, we state some of our results 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. In a *bipartite chain graph* $G = (V_s, V_p, E)$, 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 twin-free connected bipartite chain graph 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 $C_{n,n}$.

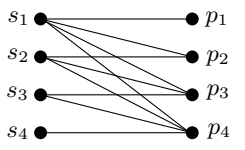


Fig. 1: The graph $C_{4,4}$

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 simple cycle with t vertices. The *domino* is the graph obtained from the cycle C_6 by adding an edge between two diametrically opposite vertices. For a graph G and a set $U \subseteq V(G)$, let $G[U]$ denote the subgraph of G induced by U.

Every bipartite graph with readability at most 1 is a disjoint union of complete bipartite graphs (also called bicliques) [6]. The characterization in the following theorem extends our understanding to graphs of readability at most 2.

Theorem 3. *A twin-free bipartite graph G has readability at most 2 iff 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.*

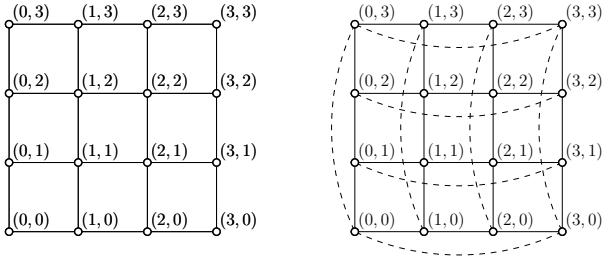


Fig. 2: The 4×4 grid $G_{4,4}$ and toroidal grid $TG_{4,4}$.

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.*

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

An efficient algorithm for readability 2 (in the full version). 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 $\mathcal{O}(2^n)$ [4]. We make some 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 their induced subgraphs (Section 5). We fully characterize the readability of grids. A *(two-dimensional) grid* is a graph $G_{m,n}$ with vertex set $\{0, 1, \ldots, m-1\} \times \{0, 1, \ldots, n-1\}$ such that there is an edge between two vertices iff the L_1-distance between them is 1. An example is shown in Figure 2. Our next theorem fully settles the question of readability of grids.

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

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

Theorem 5 has an algorithmic implication for the readability of grid graphs. A *grid graph* is an induced subgraph of a grid. Several problems are NP-hard on the class of grid graphs, including Hamiltonicity problems [12], various layout problems [8], and others (see, e.g., [7]). 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 $\{1, \ldots, k\}$ 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 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 iff the graph can be decomposed into two subgraphs G_1 and G_2 such that G_1 is a 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 the $4n \times 4n$ 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 $G = (V_s, V_p, E)$, its *twin-free reduction* $TF(G)$ is the graph with vertices being the equivalence classes of the twin relation on $V(G)$ (that is, $x \sim y$ iff x and y are twins in G), and two classes X and Y are adjacent iff $(x, y) \in E$ for some $x \in X$ and $y \in Y$. For graph theoretic terms not defined here, we refer to [24]. 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) $r(G) = r(TF(G))$.

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. 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 on the readability of twin-free bipartite chain graphs, $C_{n,n}$, and prove Theorem 1. The lower bound on their readability (Theorem 2) is proved in the full version. 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

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 $\ell(s_j)$ iff $i < j$. 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.

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

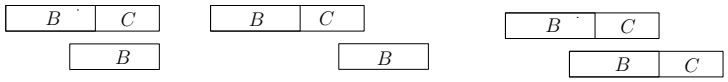
- $\forall i \in [t]$, *string* s_i *does not have a proper overlap with itself and*
- $\forall i, j \in [t]$, *string* s_i *overlaps string* s_j *iff* $i \leq j$.

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, 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 $\text{suf}_i(BC)$ 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 $\text{suf}_i(BC)$ only includes characters from C, then C overlaps A; if it includes characters from B (and the entire C) then B



(a) BC does not overlap B.

(b) BC has no proper overlap with itself.

Fig. 3: Overlaps in the proof of Lemma 2

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. □

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:

$$
\begin{array}{l|l}
r = 2 & 1\ 2 \\
r = 3 & 1\ 3\ 2 \\
r = 4 & 1\ 4\ 3\ 2 \\
r = 5 & 1\ 5\ 4\ 3\ 5\ 2 \\
r = 6 & 1\ 6\ 5\ 4\ 3\ 5\ 2 \\
r = 7 & 1\ 7\ 6\ 5\ 4\ 7\ 3\ 5\ 7\ 2
\end{array}
$$

It turns out that $|B_r|$, and, by extension, $|S_r|$, is closely related to the totient summatory function [22], 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: $\Phi(n) = \frac{3n^2}{\pi^2} + \mathcal{O}(n \log n)$ [10, 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 7) 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 8). 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 6. *If i and j are coprime then each of them is coprime with $i + j$ and 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 7. *For all $r \geq 2$, if two numbers are neighbors in B_r, 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 6, $gcd(i, r) = gcd(i, i+j) = 1$ and $gcd(r, j) = gcd(i+j, j) = 1$. Therefore, any two neighbors in B_r must be coprime. □

Claim 8. *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 6, 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}. □

4 A Characterization: Graphs with Readability at most 2

We characterize bipartite graphs with readability at most 2 by proving Theorem 3. By Lemma 1, it is enough to obtain such a characterization for connected twin-free bipartite graphs. We use this characterization (in the full version) to develop a polynomial-time algorithm for recognizing graphs of readability at most 2 and also (in Section 5) 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 two vertices at distance 3. We first define the notion of a feasible matching, which is implicitly used in the statement of Theorem 3.

Definition 3 (Feasible Matching). *A matching M in a bipartite graph G is* feasible *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.*

In the full version, we prove Theorem 3 by showing that a bipartite graph G has readability at most 2 iff G has a feasible matching. The following corollary of Theorem 3 is used in Section 5.

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. Both M_1 and M_2 are feasible matchings. Thus, by Theorem 3, G has readability at most 2. □

5 Readability of Grids and Their Induced Subgraphs

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

We now prove Theorem 5, about the readability of $G_{m,n}$. We first recall the following simple observation (which follows, e.g., from [6, Theorem 4.3]).

Lemma 5. *A bipartite graph G has: (i) $r(G) = 0$ iff G is edgeless, and (ii) $r(G) \leq 1$ iff G is P_4-free (equivalently: a disjoint union of bicliques).*

Proof (of Theorem 5). First, by Lemma 5, $r(G_{m,n})$ 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 5, their readability is 1.

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 iff G has a feasible 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 $M = \{((i, j), (i, j + 1)) \mid i \in \{0, 1\}$ and $j \in \{0, \ldots, n-2\}$ is even$\}$ is a feasible 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 4, we have that $r(G_{m,n}) \leq r(TG_{4n,4n}) \leq 3$. The proof that $r(G_{m,n}) \geq 3$ can be found in the full version. □

6 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. In the context of general bipartite graphs, it would be interesting to determine the computational complexity of determining whether the readability of a given bipartite graph is at most k, where k is either part of input or a constant greater than 2, to study the parameter from an approximation point of view, and to relate 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 result of Section 3.1 was discovered with the help of The On-Line Encyclopedia of Integer Sequences ® [22]. 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, J1-6720, and J1-7051) to M.M. The authors S.R. and N.V. were supported by NSF grant CCF-1422975 to S.R.

The author N.V. was also supported by Pennsylvania State University College of Engineering Fellowship, PSU Graduate Fellowship, and by NSF grant IIS-1453527 to P.M. The main idea of the proof of Lemma 4 was developed by V.J. in his undergraduate final project paper [13] at the University of Primorska.

References

[1] Błażewicz, J., Formanowicz, P., Kasprzak, M., Kobler, D.: On the recognition of de Bruijn graphs and their induced subgraphs. Discrete Mathematics 245(1), 81–92 (2002)
[2] Błażewicz, J., Formanowicz, P., Kasprzak, M., Schuurman, P., Woeginger, G.J.: DNA sequencing, Eulerian graphs, and the exact perfect matching problem. In: Graph-Theoretic Concepts in Computer Science. pp. 13–24. Springer (2002)
[3] Blazewicz, J., Hertz, A., Kobler, D., de Werra, D.: On some properties of DNA graphs. Discrete Applied Mathematics 98(1), 1–19 (1999)
[4] Braga, M.D.V., Meidanis, J.: 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. pp. 52–63 (2002)
[5] Chikhi, R., Jovicic, V., Kratsch, S., Medvedev, P., Milanic, M., Raskhodnikova, S., Varma, N.: Bipartite graphs of small readability. CoRR (2018), http://arxiv.org/abs/1805.04765
[6] Chikhi, R., Medvedev, P., Milanič, M., Raskhodnikova, S.: On the readability of overlap digraphs. Discrete Applied Mathematics 205, 35–44 (2016)
[7] Clark, B.N., Colbourn, C.J., Johnson, D.S.: Unit disk graphs. Discrete Mathematics 86(1-3), 165–177 (1990)
[8] Díaz, J., Penrose, M.D., Petit, J., Serna, M.J.: Approximating layout problems on random geometric graphs. J. Algorithms 39(1), 78–116 (2001)
[9] Gevezes, T.P., Pitsoulis, L.S.: Recognition of overlap graphs. Journal of Combinatorial Optimization 28(1), 25–37 (2014)
[10] Hardy, G.H., Wright, E.M.: An Introduction to the Theory of Numbers. The Clarendon Press, Oxford University Press, New York, fifth edn. (1979)
[11] Idury, R.M., Waterman, M.S.: A new algorithm for DNA sequence assembly. Journal of Computational Biology 2(2), 291–306 (1995)
[12] Itai, A., Papadimitriou, C.H., Szwarcfiter, J.L.: Hamilton paths in grid graphs. SIAM J. Comput. 11(4), 676–686 (1982)
[13] Jovičić, V.: Readability of digraphs and bipartite graphs (2016), final project paper. University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies, Koper, Slovenia, 2016. Available at https://arxiv.org/abs/1612.07113
[14] Kasprzak, M.: Classification of de Bruijn-based labeled digraphs. Discrete Applied Mathematics (2016)
[15] Li, X., Zhang, H.: Characterizations for some types of DNA graphs. Journal of Mathematical Chemistry 42(1), 65–79 (2007)
[16] Li, X., Zhang, H.: Embedding on alphabet overlap digraphs. Journal of Mathematical Chemistry 47(1), 62–71 (2010)
[17] Miller, J.R., Koren, S., Sutton, G.: Assembly algorithms for next-generation sequencing data. Genomics 95(6), 315–327 (2010)
[18] Myers, E.W.: The fragment assembly string graph. In: ECCB/JBI. p. 85 (2005)

[19] Nagarajan, N., Pop, M.: Sequence assembly demystified. Nature Reviews Genetics 14(3), 157–167 (2013)
[20] Pendavingh, R., Schuurman, P., Woeginger, G.J.: Recognizing DNA graphs is difficult. Discrete Applied Mathematics 127(1), 85–94 (2003)
[21] Simpson, J.T., Durbin, R.: Efficient de novo assembly of large genomes using compressed data structures. Genome Research (2011)
[22] Sloane, N.J.A.: The On-Line Encyclopedia of Integer Sequences (2016), published electronically at `https://oeis.org`
[23] Tarhio, J., Ukkonen, E.: A greedy approximation algorithm for constructing shortest common superstrings. Theoretical Computer Science 57(1), 131–145 (1988)
[24] West, D.B.: Introduction to Graph Theory. Prentice Hall, Inc., Upper Saddle River, NJ (1996)