Decentralized Graph Processing for Reachability Queries

Joël Mathys$^{1[0000-0003-1601-5234]}$, Robin Fritsch^{1}, and Roger Wattenhofer$^{1[0000-0002-6339-3134]}$

ETH Zürich, Switzerland
{jmathys, rfritsch, wattenhofer}@ethz.ch

Abstract. Answering queries on large graphs is an essential part of data processing. In this paper, we focus on determining reachability between vertices. We propose a labeling scheme which is inherently distributed and can be processed in parallel. We study what properties make it difficult to find a good reachability labeling scheme for directed graphs. We focus on the genus of a graph. For graphs of bounded genus g, we design a labeling scheme of length $\mathcal{O}(g \log n + \log^2 n)$. We also prove that no labeling schemes with labels shorter than $\Omega(\sqrt{g})$ exist for this graph class.

Keywords: Reachability query · Labeling scheme · Data-mining-ready structures and pre-processing

1 Introduction

Databases and web services employ some variant of precomputation in order to answer queries reasonably quickly. Several applications, such as XML parsing, logical reasoning on semantic web RDF/OWL data, querying protein-protein interaction networks, lineage tracking for scientific workflows or routing work directly on graph data [28, 10, 30, 18]. Determining the relationship between nodes in such graphs is a fundamental operation to efficiently answer queries. In this paper, we study the classical problem of determining reachability in graphs: Given a directed graph G, we want to answer whether a node u can reach a node v, for arbitrary u, v. Furthermore, it might be to inefficient to answer such reachability queries by running a naive linear time algorithm. If the graph is large, we can also not store all possible answers. Instead, we propose to construct a reachability oracle in the form of a labeling scheme: We store some additional information (a label) with each node. When we want to answer a reachability query for nodes u, v, we simply look up the labels of nodes u and v, and deduce the reachability just from these labels. Such an approach can also deal with graphs that are stored in a distributed fashion, since the labels of u and v do not need to be stored on the same machine. Furthermore, the labeling scheme is highly parallelizable as we only require read access to the labels of both nodes.

In the literature, such a distributed oracle is known as an informative labeling scheme [23]. A labeling scheme consists of an encoder l and a decoder d. The encoder l assigns a *label* (a bitstring) $l(u)$ to each vertex u of the graph. The decoder d takes the labels $l(u)$ and $l(v)$ as input and then directly decides if u can reach v in G, without using any other information.

The main objective when constructing a labeling scheme is to minimize the maximum label length assigned to any vertex. But how difficult is this task? For which graphs do we get a reasonable label length? Or more precisely, can we leverage the intrinsic properties of the given data to model graphs for which we can create short labeling schemes?

A straightforward idea is to use the sparsity and locality in the given data. The graph representation often only contains a small fraction of all possible connections. A first attempt is to consider that each element in the data is only directly related to a few other elements. Such an assumption translates naturally to the graph domain as they can be modeled with bounded degree graphs. However, such a local restriction is not strong enough. We show that even for sparse graphs that have a bounded outdegree Δ it is not possible to construct short schemes. We prove a lower bound of $\Omega(\sqrt{n\Delta})$, so already graphs with outdegree 2 require labels of length at least $\Omega(\sqrt{n})$.

In search for a more suitable characterization of the graphs, we consider a more global parameter, the genus. Graphs of bounded genus are a natural generalization of planar graphs, since planar graphs can be characterized as graphs of genus 0. For planar graphs, in a seminal paper, Thorup [27] constructed a labeling scheme of length $\mathcal{O}(\log^2 n)$. In the same paper, he posed as an open question whether this approach could be extended to graph classes excluding minors and bounded genus. Building upon Thorup's work, as well as on the work of Gilbert et al. [17] on separator sets in bounded genus graphs and the work of Kawarabayashi et al. [21] on labeling schemes for distances, we construct a new labeling scheme of length $\mathcal{O}(g \log n + \log^2 n)$. Furthermore, we provide a lower bound that applies to any labeling scheme designed for the class of bounded genus graphs of $\Omega(\sqrt{g})$, even if the genus is subquadratic in the size of the graph. This means that the genus of a graph is a possible indicator on the complexity of devising short labelings.

2 Related Work

Labeling schemes have been around for more than 50 years, making an appearance in the context of information theory by Breuer [9]. Ever since, labeling schemes have been studied for several different types of queries. These include, but are not limited to, labeling schemes for adjacency [20, 2, 6, 8], lowest common ancestor [5, 23] and distance [1, 27, 16, 21]. All known to us online maps use distance labeling to compute shortest paths.

Labeling schemes relating to reachability in directed graphs, sometimes also referred to as ancestry queries, have been studied extensively on trees as they can be used for improving the performance of XML search engines [15, 7]. The problem was first introduced by Kannan et al. [20], who presented a $\mathcal{O}(\log n)$ labeling scheme for reachability on trees. Schemes were improved [7, 13, 15] to match the lower bound of $\log n + \Omega(\log \log n)$ derived by Alstrup et al. [4]. There have also been publications on constructing labeling schemes for reachability for the class of planar graphs as they are very useful to answer routing queries in online maps. Working on distance oracles for approximate distances, Thorup [27] constructed a labeling scheme for reachability for directed planar graphs of length $\mathcal{O}(\log^2 n)$. His work also applied to labeling schemes of approximate distances for undirected planar graphs. Kawarabayashi et al. [21] then generalized the labeling scheme of Thorup for approximate distances in undirected graphs to bounded genus graphs. It uses a tree-cotree decomposition introduced by Eppstein [14] to reduce the problem to the planar case. More recently, there has been work on labeling schemes for other classes of directed graph, such as treewidth [19] or compressing the transitive closure [8]. But the relevance of reachability goes beyond trees and planar graphs. For instance, reachability labels are used for efficient lineage tracking [18].

There is an extensive list of publications, which focus on building implemented systems to answer reachability queries. Several approaches build on the ideas of chain covers [25], tree covers [3], hop labeling [11] or a combination of these techniques [28, 10]. On the other hand, some of these systems allow for additional computation on the graph at querying time. These approaches are usually based on interval labeling [30, 24], sampling linear extensions [12] or set containment [29]. Whenever the information given by the precomputation is not sufficient, they fall back to a linear search to answer the query. We refer to [29] for an in-depth coverage of existing work.

This paper analyses what parameters and restrictions are required for graph classes in order to construct short and efficient reachability schemes. These insights are important to construct new schemes and can further be used to estimate theoretical limitations of practical systems.

Labeling Schemes for Reachability		
Graph Class	Lower Bound	Upper Bound
general graphs	$\Omega(n)$ [8]	$\mathcal{O}(n)$ [8]
trees	$\Omega(\log n)$	$\mathcal{O}(\log n)$ [13, 15]
outdegree $\Delta \geq 2$	$\boldsymbol{\Omega(\sqrt{n\Delta})}$	$\mathcal{O}(n)$ [8]
planar	$\Omega(\log n)$	$\mathcal{O}(\log^2 n)$ [27]
genus g	$\boldsymbol{\Omega(\sqrt{g})}$	$\boldsymbol{\mathcal{O}(g \log n + \log^2 n)}$

Table 1. An overview of the asymptotic length of reachability labeling schemes for different graph classes. The bounds printed in bold are shown in this paper. (The $\Omega(\log n)$ lower bounds are straightforward, since a labeling scheme for paths requires at least $\Omega(\log n)$ bits.)

Table 1 shows known lower and upper bounds on the length of reachability labeling schemes for a number of graph classes. The bounds we prove in this paper are printed in bold.

In this paper, we use the ideas introduced by Kawarabayashi et al. [21] to construct a labeling scheme for reachability in directed graphs of bounded genus. However, instead of relying on the decomposition of Eppstein [14], we adapt the work of Gilbert et al. [17] on balanced vertex separators in bounded genus graphs. Gavoille et al. [16] proved that labeling schemes for exact distances in undirected graphs of bounded degree must be of length at least $\Omega(\sqrt{n})$. In our work, we show that this bound can be adapted to the case of reachability in directed graphs of bounded degree.

3 Graphs of Bounded Degree

When processing data represented as graphs, the data usually follows a given structure. Meaning that the number of edges is not quadratic, but roughly linear in the number of vertices. Therefore, we deal with very sparse graphs. However, is it enough to build a short labeling scheme by considering a sparsity constraint? We study a natural way to enforce sparsity by locally enforcing a maximum degree per vertex.

While restricting the degree of vertices of the graph seems like an appropriate choice to limit a graph's complexity, it turns out not to be the right parameter to explain the difficulty of constructing short labeling schemes for reachability. In the following, we prove a lower bound of $\Omega(\sqrt{n\Delta})$ on the length of a reachability labeling scheme for graphs with degree $\Delta \geq 2$. In order to achieve this, we introduce a transformation Ψ_Δ.

Note that if the outdegree is restricted to be 1, we can apply the labeling scheme for trees. Recall that for reachability, we always work with directed graphs. For a vertex v we denote its outdegree with $\deg^+(v)$ and its indegree with $\deg^-(v)$. Note that it does not matter wheter we restrict either the in- or outdegree whenever we are interested in a labeling scheme for reachability. In fact, they are interchangeable. We can reverse the directions of all edges of a digraph G to swap in- and outdegree and obtain G'. We then have that $u \rightsquigarrow_G v$ (u reaches v in G) if and only if $v \rightsquigarrow_{G'} u$. Therefore, we usually only restrict the outdegree of a graph.

3.1 Degree Graph Transformation

The idea behind the transformation is that any digraph can be transformed into another digraph with smaller outdegree, while preserving the reachability relation. We replace each original vertex of the digraph with multiple vertices of smaller outdegree linked together by a chain. More formally, the transformation Ψ_Δ takes an arbitrary digraph G and transforms it into a digraph G' with maximum outdegree Δ.

Note that Gavoille et al. [16] used a similar technique to derive a $\Omega(\sqrt{n})$ lower bound for determining exact distances using labeling schemes in undirected bounded degree graphs. The lower bound technique and in particular the Ψ_Δ transformation can be applied to directed graphs and were found independently.

Recall that for general digraphs, there is a lower bound linear in the size of the graph.

Theorem 1 (Digraph Lower Bound [8]). *Let $\mathcal{G}$ be the class of digraphs on n vertices. Every labeling scheme for reachability for the class $\mathcal{G}$ has length at least $L = \frac{n}{4} = \Omega(n)$.*

Lemma 1. *Let G be an arbitrary digraph on n vertices with maximum outdegree ω. The transformation Ψ_Δ constructs a digraph $G' = \Psi_\Delta(G)$ with the following properties.*

1. *(Bounded Degree) G' has maximum outdegree $\Delta \geq 2$.*
2. *(Additional Vertices) G' has at most $n\lceil \frac{\omega}{\Delta - 1} \rceil$ vertices.*
3. *(Reachability) $V(G) \subseteq V(G')$, furthermore for any $u, v \in V(G)$: u can reach v in G if and only if u can reach v in G'.*

Proof. We only outline the construction of the transformation here and refer to the full version of a paper for a more detailed construction and formal proof of the properties.

The transformation applies to each vertex independently. More specifically, in G' each vertex v of G is replaced by a chain of new vertices. Among these new vertices the outgoing edges are distributed in a way that ensures that all vertices in G' have outdegree at most Δ.

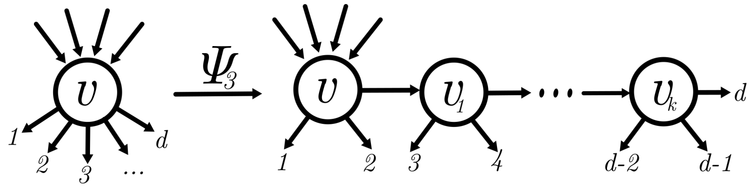

Fig. 1. The Ψ_Δ transformation applied to a vertex v of outdegree d with $\Delta = 3$. On the right side, $v = v_0$ and its virtual vertices $v_1, ..., v_k$ are displayed.

Note that the shape with which we replace each vertex with could be changed to resemble a tree instead of a chain. However, it is more convenient to have it be a simple path. Furthermore, we could also transform the indegree of the graph using the same technique. Finally, we need to relate the Ψ_Δ transformation to the labeling scheme.

3.2 Bounded Degree Lower Bound

We can now use the transformation Ψ_Δ to get a lower bound for bounded degree graphs. The main idea behind the lower bound is that if there were a shorter labeling scheme, we could use it to construct a shorter scheme for general digraphs.

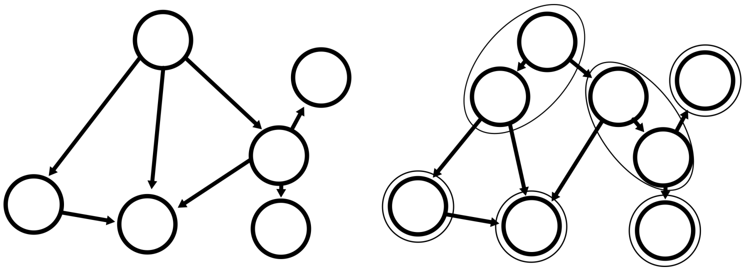

Fig. 2. On the left side a graph G that contains nodes with degree larger than Δ. On the right side the Ψ_Δ transformed graph G' with maximum outdegree $\Delta = 2$. The reachability information of G is preserved in G'.

Theorem 2 (Bounded Degree Lower Bound). *Let $\mathcal{G}_\Delta$ be the class of digraphs on n vertices with outdegree $\Delta \geq 2$. Every labeling scheme for reachability for the class $\mathcal{G}_\Delta$ has length at least $L = \Omega(\sqrt{n\Delta})$.*

Proof. Assume for the sake of contradiction that there exists a labeling scheme for reachability for the class $\mathcal{G}_\Delta$ and which has length $L = o(\sqrt{n\Delta})$. We now apply this scheme to $\mathcal{G}$, the class of digraphs on n vertices. Now take any digraph $G \in \mathcal{G}$ and Ψ_Δ transform G to the graph $G' \in \mathcal{G}'$. Note that due to the reachability property of the transformation any labeling for $G' = \Psi_\Delta(G)$ is also a valid labeling for G. Furthermore, G' now contains at most $n \cdot \lceil \frac{n}{\Delta-1} \rceil$ vertices and every vertex has outdegree at most Δ by the properties of the Ψ_Δ transformation.

We now apply the labeling scheme of length $o(\sqrt{n\Delta})$ to $\mathcal{G}'$, which also is a labeling scheme for $\mathcal{G}$. Note that due to $\Delta \geq 2$ the term $\lceil \frac{n}{\Delta-1} \rceil$ is at most $\frac{2n}{\Delta}$.

$$L = o\left(\sqrt{\Delta |V(G')|}\right) \leq o\left(\sqrt{\Delta n \cdot \left\lceil \frac{n}{\Delta - 1} \right\rceil}\right) \leq o\left(\sqrt{\Delta n \cdot \frac{2n}{\Delta}}\right) = o\left(\sqrt{n^2}\right) = o(n) \tag{1}$$

However, we already know by Theorem 1 that any labeling scheme for $\mathcal{G}$ must use at least $\Omega(n)$ bits, a contradiction. Therefore, any labeling scheme for reachability for the class $\mathcal{G}_\Delta$ has length at least $L = \Omega(\sqrt{n\Delta})$.

4 Graphs of Bounded Genus

After studying the maximum degree of a graph, we now turn to another parameter: the genus of a graph. A graph of genus g can be embedded in a surface of genus g so that none of its edges are crossing. Graphs of bounded genus are a natural extension of planar graphs, since the latter are exactly the graphs of genus 0. We already know that low genus graphs, such as planar graphs, have very short labeling schemes. On the other hand, general directed graphs with many edges have a lot of crossings. In the following, we show that the genus of a graph is a useful indicator of how difficult it is to construct a short labeling.

Building upon the previous work of Thorup [27] on labeling schemes in planar graphs, the work of Gilbert et al. [17] on separator sets in bounded genus graphs and the work of Kawarabayashi et al. [21] on labeling schemes for distances, we construct a new labeling scheme of length $\mathcal{O}(g \log n + \log^2 n)$. Furthermore, we provide a lower bound on the length of any labeling scheme of $\Omega(\sqrt{g})$ for the class of bounded genus graphs. Moreover, the provided proof extends to graphs whose genus is linear in the size of the graph.

First, we recall the layering technique introduced by Thorup, as our construction heavily relies on it. We proceed with a slight variation of the separator proof of Gilbert et al. to reduce a given graph of bounded genus to a planar graph. Then we combine the tools to construct a new labeling scheme for bounded genus graphs. Before we begin, we briefly discuss a frequently used method (e.g. in [27]) that allows us to remove directed paths from the graphs.

4.1 Preliminaries

In the following, we describe two methods that we use as building blocks in our construction. The first method is used to handle a directed path p. It allows us to reduce the construction of the labeling to the graph without p by storing an additional $\mathcal{O}(\log n)$ bits per vertex.

Lemma 2 (Path Labeling [27]). *Let G be a digraph on n vertices, P a set of t directed paths in G and a labeling l' for the graph $G \setminus V(P)$. By storing $\mathcal{O}(t \log n)$ additional bits per vertex, we can extend the labeling l' to a labeling l for G.*

We leave the rigorous proof itself for the full version. The idea is that each vertex stores the first vertex it can reach on a path p and the last vertex on p that can reach it. Given two vertices we can then decide if there is a path between the two vertices that uses some vertex in p.

The second method, introduced by Thorup [27], reduces the problem of constructing a reachability scheme for the whole graph to a sequence of smaller graphs that are more local. By more local, we mean that it is possible that in the original graph, many parts of the graph will never interact with each other as there are no directed paths between them.

Definition 1 (t-layered Spanning Tree [27]). *A t-layered spanning tree T in a digraph G is a disoriented rooted spanning tree such that any path in T starting from the root is the concatenation of at most t directed paths in G.*

Lemma 3 (Digraph Layering [27]). *Given an arbitrary digraph G, there exists a series of digraphs $G_0, ..., G_{k-1}$ with $k \leq n$ so that*

1. *(Reachability) Each vertex v has an index $\tau(v)$, so that a vertex w is reachable from v in G if and only if w is reachable from v in $G_{\tau(v)}$ or $G_{\tau(v)-1}$.*
2. *(Spanning Tree) Each $G_i = (V_i, E_i)$ is a digraph with a 2-layered spanning tree T_i rooted at r_i.*

3. *(Minor closed) G_i is a minor of G. If G has genus at most g, then G_i also has genus at most g.*
4. *(Linear Size) The total number of edges and vertices over all G_i is linear in the number of edges and vertices in G. Furthermore, every vertex v appears as a non-root vertex only in $G_{\tau(v)}$ and $G_{\tau(v)-1}$.*

We now relate the digraph layering construction to labeling schemes. Recall that each vertex v only appears in two digraphs $G_{\tau(v)-1}, G_{\tau(v)}$. Furthermore, every vertex w which is reachable from v is also reachable from v in one of the two digraphs. The idea is to construct a labeling for each digraph separately. Then to obtain the labeling for the original graph, we concatenate the sublabelings of the two graphs $G_{\tau(v)-1}, G_{\tau(v)}$. This way, the length of the labeling is at most twice the length of the sublabelings. We refer to the full version of the paper for the proof.

Lemma 4 (Digraph Layering Labeling [27]). *Given an arbitrary digraph G, its sequence of digraphs $G_0, ..., G_{k-1}$ obtained by digraph layering and a labeling l_i for each digraph G_i. We can construct a labeling l for G. Furthermore, l is of length at most $2 \log n + 2 \max |l_i|$.*

4.2 Planarizing Technique

In this section, we are given a bounded genus graph, and we want to reduce its genus to make it planar. The idea is to relate the problem of finding a reachability scheme for bounded genus graphs to the labeling scheme for planar graphs.

Given a digraph and a spanning tree, we construct a set of undirected paths so that the removal of these paths makes the graph planar. Moreover, if the tree is a 2-layered spanning tree, we can guarantee that the set of paths consists of not too many directed paths.

Given an arbitrary digraph G of genus g with a spanning tree T, we want to reduce the genus of G by removing certain paths of T. More formally, we are given T and build a set P consisting of paths of T starting at the root. The goal is that $G \setminus V(P)$ is a planar graph, not necessarily connected. Furthermore, if the given spanning tree T is a k-layered spanning tree, the set P consists of at most $\mathcal{O}(kg)$ directed paths.

Lemma 5 (Planarizing Tree). *Given a digraph G of genus at most g and a spanning tree T with root r, there exists a set of paths P so that*

1. *(Planarizing) Each connected component of $G \setminus V(P)$ is planar.*
2. *(Pathset) P consists of at most $\mathcal{O}(g)$ undirected paths in T starting at the root vertex r.*
3. *(Layered Tree) If T is a k-layered spanning tree, then P consists of at most $\mathcal{O}(kg)$ directed paths of T.*

We only outline the construction of the set P and refer to [17] for a more in-depth coverage. The main difference between the original proof of Gilbert et al. and our proof is the choice of the spanning tree T. In the original proof the spanning tree had to be of small depth to bound the lengths (and number of vertices) of the undirected paths of P. This leads to a small vertex separator set. However, we only require that the set consists of a small number of paths which might consist of many vertices. We can therefore take an arbitrary spanning tree T as the origin of the construction instead of a low depth tree.

Instead of directly constructing the set P in the graph G, we remove parts of the graph that do not belong to P. The part that remains after removing certain edges and vertices is the desired set P. Everything that we remove in this process, namely $G \setminus V(P)$, is planar.

First, the graph is embedded in a surface of genus g with a spanning tree T. Then, non-tree edges are repeatedly removed to reduce the number of faces while preserving the embedding. Afterwards, all vertices of degree one are repeatedly removed until none are left. What remains is a graph G' spanned by $T' \subseteq T$. Due to Euler's Theorem ($v - e + f = 2 - 2g$) the number of non-tree edges in G' and as a consequence also the number of leaves of T can be upper bounded by $\mathcal{O}(g)$. Thus, P can be covered by $\mathcal{O}(g)$ undirected paths. In particular, if T is a k-layered spanning tree P consists of $\mathcal{O}(kg)$ directed paths.

4.3 Bounded Genus Labeling Scheme

Now that we have introduced the necessary technical tools, we put them together to construct the labeling scheme for bounded genus graphs.

We remark that the labeling scheme of Kawarabayashi et al. [21] for distances in bounded genus graphs builds on the same idea. However, for reachability we work with directed graphs. Therefore, the technique is not directly applicable. Furthermore, instead of using the more complex tree-cotree decomposition of Eppstein [14], we rely on the separator construction of Gilbert et al. [17] to planarize the graph.

The idea of the construction is as follows. First, we construct a sequence of digraphs according to the digraph layering. Due to the construction, the genus does not increase for these layered digraphs. We then use the planarizing technique from the previous section to reduce them to planar graphs. Finally, we can apply the scheme devised by Thorup [27] for the planar remainder of the graph.

Theorem 3 (Bounded Genus Labeling). *Let $\mathcal{G}$ be the class of digraphs on n vertices with genus at most g. There exists a labeling scheme for reachability for the class $\mathcal{G}$ of length $L = \mathcal{O}(g \log n + \log^2 n)$.*

Proof. Let $G \in \mathcal{G}$ be an arbitrary digraph on n vertices of genus at most g. First we construct a sequence of layered digraphs $G_0, ..., G_{k-1}$ according to Lemma 3. Due to Lemma 4 it is sufficient for us to construct a labeling scheme of length $\mathcal{O}(g \log n + \log^2 n)$ for each G_i separately.

Recall that each G_i has a 2-layered spanning tree T_i rooted at r_i. Furthermore, G_i is a minor of G, therefore the genus of G_i is at most g.

Let H be an arbitrary G_i of the layered digraph construction. We can apply Lemma 5 on H using the 2-layered spanning tree. This gives us a set P of at most $2 \cdot 4g = \mathcal{O}(g)$ directed paths whose removal yields a planar graph H'.

Using Lemma 2, we can store $\mathcal{O}(g \log n)$ bits for the paths in P and reduce the problem to constructing a labeling scheme for $H \setminus V(P) = H'$. As the remaining graph H' is planar, we can apply the labeling scheme of Thorup [27] to construct a labeling of length $\mathcal{O}(\log^2 n)$.

Therefore, we have constructed a labeling scheme for the graph G using a total of $\mathcal{O}(g \log n + \log^2 n)$ bits.

4.4 Bounded Genus Lower Bound

In this section, we want to prove a lower bound for the length of any labeling scheme for reachability in bounded genus graphs. We do this by relating the class of bounded degree graphs to the class of bounded genus graphs. We first need two technical lemmas to relate the genus to the number of edges in a graph and to get a bound on the genus of the complete graph.

Lemma 6 (Genus Upper Bound). *Let G be an arbitrary graph on v vertices, $e \geq 1$ edges and genus g. Then the genus is at most the number of edges: $g \leq e$.*

Proof. Let G be a graph of genus g with v vertices and $e \geq 1$ edges. Let f be the number of faces of an embedding of G in a surface of genus g. Then, according to Euler's formula

$$v - e + f = 2 - 2g. \tag{2}$$

Rearranging the terms gives the desired bound on the genus. Note that v, e and f are all nonnegative integers.

$$g = \frac{2 - v + e - f}{2} \leq 1 + \frac{e}{2} \leq e \tag{3}$$

Lemma 7 (Genus of complete graph [22]). *The complete graph K_g has genus $\Theta(g^2)$ and as a consequence every graph with a K_g minor must have genus at least $\lceil \frac{(g-4)(g-5)}{12} \rceil = \Theta(g^2)$.*

Now we can to relate the lower bound of the bounded degree graphs to the bounded genus graphs.

Theorem 4 (Bounded Genus Labeling Lower Bound). *Let $\mathcal{G}$ be the class of digraphs on n vertices of genus at most g. Every labeling scheme for reachability for the class $\mathcal{G}$ has length at least $L = \Omega(\sqrt{g})$.*

Proof. Assume for the sake of contradiction that there exists a labeling scheme for reachability for the class $\mathcal{G}$ and has length $L = o(\sqrt{g})$. The idea is to take this scheme and apply it to the class of bounded degree graphs $\mathcal{G}_\Delta$. More formally, $\mathcal{G}_\Delta$ is the class of digraphs on n vertices with outdegree at most Δ. These graphs have at most $n \cdot \Delta$ edges. By Lemma 6, their genus is bounded by $n \cdot \Delta$. Now we use the labeling scheme for bounded genus graphs of length $o(\sqrt{g})$ for $\mathcal{G}_\Delta$.

$$L = o(\sqrt{g}) \leq o(\sqrt{n\Delta}) \tag{4}$$

However, we already know by Theorem 2 that any labeling scheme for $\mathcal{G}_\Delta$ must use at least $\Omega(\sqrt{n\Delta})$ bits, a contradiction. Therefore, any labeling scheme for reachability for the class $\mathcal{G}$ has length at least $L = \Omega(\sqrt{g})$.

Note that the $\Omega(\sqrt{g})$ bound can be derived by arguing that the genus is always at most the number of vertices squared. However, the proof presented here is stronger in the following sense. The simple proof only shows the result for the class of all graphs, which includes the graphs of genus of order up to $\Theta(n^2)$. However, the proof can't be applied if the class of graphs is more restricted, i.e. graphs of genus $\Theta(n^2)$ are not included. Whereas with our proof, the result also follows for more general classes of graphs including the graphs of genus linear in the size of the graph.

Note that we could have also derived the square root bound by arguing that the genus is always at most the number of vertices squared. However, we show a specific class of graphs with genus linear in the size of the graph to which the square root bound applies, instead of only graphs with genus of order $\Theta(n^2)$.

Finally, we study that the lower bound for bounded genus graphs can be transferred to the class of minor excluded graphs. We say that a graph G contains a minor H if H can be obtained through a series of edge contractions, vertex deletions, or edge deletions. The class of minor excluded graphs consists of the graphs which exclude all graphs H of a family $\mathcal{H}$ as a minor. This is of interest as many classes of graphs can be characterized through excluding minors. In particular, the class of planar graphs can be defined as the graphs which exclude $K_{3,3}$ and K_5, known as Kuratowski's Theorem [26]. Note that for reachability, we work with directed graphs. In the context of minor excluded graphs, we work with the undirected graph underlying G whenever necessary.

Theorem 5 (Minor Excluded Lower Bound). *Let $\mathcal{G}_H$ be the class of digraphs on n vertices, which excludes the graph H on h vertices as a minor. Every labeling scheme for reachability for the class $\mathcal{G}_H$ has length at least $L = \Omega(h)$.*

Proof. First, let us consider the class $\mathcal{G}_{K_h}$ which excludes the complete graph on h vertices as a minor. Note that $\mathcal{G}_{K_h}$ is a subclass of $\mathcal{G}_H$ as any graph that excludes H as a minor can not have K_h as a minor. Therefore, any lower bound we derive for $\mathcal{G}_{K_h}$ must also apply to $\mathcal{G}_H$.

Assume for the sake of contradiction that there exists a labeling scheme for the class $\mathcal{G}_{K_h}$ of length $L = o(h)$. We will now take this scheme and apply it to the class of bounded genus graphs $\mathcal{G}$ of genus at most g. Recall that due to Lemma 6, the genus does not increase for any minor. Furthermore, the complete graph $K_{g'}$ where $g' = \lceil 12(\sqrt{g+1}+5) \rceil$ has genus at least $g+1$ due to Lemma 7. As a consequence, any graph with a $K_{g'}$ minor has genus at least $g+1$. Therefore,

$\mathcal{G}$ excludes $K_{g'}$ and we can apply the labeling scheme for excluded graph minors of length $o(h)$ on $\mathcal{G}$.

$$L = o(h) = o(g') \leq o\left(\lceil 12(\sqrt{g+1}+5)\rceil\right) = o\left(\sqrt{g}\right) \qquad (5)$$

However, we already know by Theorem 4 that any labeling scheme for $\mathcal{G}$ must use at least $\Omega(\sqrt{g})$ bits, a contradiction. Therefore, any labeling scheme for reachability for the class $\mathcal{G}_{K_h}$ and $\mathcal{G}_H$ has length at least $L = \Omega(h)$.

In this section we have strengthened existing results and showed how minor excluded, bounded genus and bounded degree graphs are intertwined. We view this as a first step to obtain tighter lower bounds, which might be harder to obtain and require new insights.

Finally, a short remark on the tightness of our results. For undirected graphs which exclude H as a minor there is a property called path separability studied by Abraham et al. [1]. How path separability behaves for the directed case is not yet as well understood. If such graphs were $\mathcal{O}(h)$ path separable, this would further imply the existence of a labeling scheme of length $\mathcal{O}(h\log^2(n))$ which could also be applied to graphs of bounded genus and graphs of bounded degree. In this case, the derived bounds would be tight up to $\log(n)$ factors.

5 Conclusion

Processing graph data is essential for many applications. Furthermore, determining the relationship between vertices quickly and efficiently is central to get good performance. We study how to answer reachability queries using a labeling scheme. A labeling scheme can be stored in a distributed fashion by design as each node only needs to store his own label. Therefore, they are well suited to be applied to large graphs. Furthermore, they allow for parallel processing as only read operations on the labels are needed. However, a key challenge is how to design short labeling schemes and determine whether it is even possible.

In order to determine an appropriate parameter to characterize the difficulty of designing short labeling schemes for reachability we study the degree and genus of a graph as a measure. We first study sparse graphs, where we enforce local sparseness through limiting the maximum degree. It turns out, that even for constant degree graphs there are no short schemes as we prove a $\Omega(\sqrt{n\Delta})$ lower bound for graphs of outdegree Δ. On the other hand, we present a novel labeling scheme for graphs of bounded genus of length $\mathcal{O}(g\log n + \log^2 n)$.

Furthermore, any labeling scheme for graphs of bounded genus must use at least $\Omega(\sqrt{g})$ bits, even if the genus is subquadratic in the size of the graph. This means that the genus of a graph is a possible indicator on the complexity of devising short labelings. Moreover, the result can be generalized to minor excluded graphs. However, it remains an open question whether there exists a polylogarithmic labeling scheme. Similar results for undirected graphs [1] suggest that this might be possible. However, adapting existing results to the directed case has not been achieved so far and might need novel insights.

References

1. Abraham, I., Gavoille, C.: Object location using path separators. In: Proceedings of the Twenty-Fifth Annual ACM Symposium on Principles of Distributed Computing. p. 188–197. PODC '06, Association for Computing Machinery, New York, NY, USA (2006). https://doi.org/10.1145/1146381.1146411, https://doi.org/10.1145/1146381.1146411
2. Adjiashvili, D., Rotbart, N.: Labeling schemes for bounded degree graphs. In: Esparza, J., Fraigniaud, P., Husfeldt, T., Koutsoupias, E. (eds.) Automata, Languages, and Programming. pp. 375–386. Springer Berlin Heidelberg, Berlin, Heidelberg (2014)
3. Agrawal, R., Borgida, A., Jagadish, H.V.: Efficient management of transitive relationships in large data and knowledge bases. In: Proceedings of the 1989 ACM SIGMOD International Conference on Management of Data. p. 253–262. SIGMOD '89, Association for Computing Machinery, New York, NY, USA (1989). https://doi.org/10.1145/67544.66950, https://doi.org/10.1145/67544.66950
4. Alstrup, S., Bille, P., Rauhe, T.: Labeling schemes for small distances in trees. SIAM Journal on Discrete Mathematics **19**(2), 448–462 (2005). https://doi.org/10.1137/S0895480103433409, https://doi.org/10.1137/S0895480103433409
5. Alstrup, S., Gavoille, C., Kaplan, H., Rauhe, T.: Nearest common ancestors: A survey and a new algorithm for a distributed environment. Theory of Computing Systems **37**(3), 441–456 (May 2004). https://doi.org/10.1007/s00224-004-1155-5, https://doi.org/10.1007/s00224-004-1155-5
6. Alstrup, S., Kaplan, H., Thorup, M., Zwick, U.: Adjacency labeling schemes and induced-universal graphs. In: Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing. p. 625–634. STOC '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2746539.2746545, https://doi.org/10.1145/2746539.2746545
7. Alstrup, S., Rauhe, T.: Improved labeling scheme for ancestor queries. In: Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms. p. 947–953. SODA '02, Society for Industrial and Applied Mathematics, USA (2002)
8. Bonamy, M., Esperet, L., Groenland, C., Scott, A.: Optimal Labelling Schemes for Adjacency, Comparability, and Reachability, p. 1109–1117. Association for Computing Machinery, New York, NY, USA (2021), https://doi.org/10.1145/3406325.3451102
9. Breuer, M.: Coding the vertexes of a graph. IEEE Trans. Inf. Theory **12**, 148–153 (1966)
10. Cheng, J., Huang, S., Wu, H., Fu, A.W.C.: Tf-label: A topological-folding labeling scheme for reachability querying in a large graph. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. p. 193–204. SIGMOD '13, Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2463676.2465286, https://doi.org/10.1145/2463676.2465286
11. Cohen, E., Halperin, E., Kaplan, H., Zwick, U.: Reachability and distance queries via 2-hop labels. SIAM Journal on Computing **32**(5), 1338–1355 (2003). https://doi.org/10.1137/S0097539702403098, https://doi.org/10.1137/S0097539702403098

12. da Silva, R.F., Urrutia, S., Hvattum, L.M.: Extended high dimensional indexing approach for reachability queries on very large graphs. Expert Systems with Applications **181**, 114962 (2021). https://doi.org/https://doi.org/10.1016/j.eswa.2021.114962, https://www.sciencedirect.com/science/article/pii/S0957417421004036
13. Dahlgaard, S., Knudsen, M.B.T., Rotbart, N.: A simple and optimal ancestry labeling scheme for trees. In: Halldórsson, M.M., Iwama, K., Kobayashi, N., Speckmann, B. (eds.) Automata, Languages, and Programming. pp. 564–574. Springer Berlin Heidelberg, Berlin, Heidelberg (2015)
14. Eppstein, D.: Dynamic generators of topologically embedded graphs. CoRR **cs.DS/0207082** (2002), https://arxiv.org/abs/cs/0207082
15. Fraigniaud, P., Korman, A.: An optimal ancestry labeling scheme with applications to xml trees and universal posets. J. ACM **63**(1) (Feb 2016). https://doi.org/10.1145/2794076, https://doi.org/10.1145/2794076
16. Gavoille, C., Peleg, D., Pérennes, S., Raz, R.: Distance labeling in graphs. In: Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms. p. 210–219. SODA '01, Society for Industrial and Applied Mathematics, USA (2001)
17. Gilbert, J.R., Hutchinson, J.P., Tarjan, R.E.: A separator theorem for graphs of bounded genus. Journal of Algorithms **5**(3), 391–407 (1984). https://doi.org/https://doi.org/10.1016/0196-6774(84)90019-1, https://www.sciencedirect.com/science/article/pii/0196677484900191
18. Heinis, T., Alonso, G.: Efficient lineage tracking for scientific workflows. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. p. 1007–1018. SIGMOD '08, Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1376616.1376716, https://doi.org/10.1145/1376616.1376716
19. Jain, R., Tewari, R.: Reachability in High Treewidth Graphs. In: Lu, P., Zhang, G. (eds.) 30th International Symposium on Algorithms and Computation (ISAAC 2019). Leibniz International Proceedings in Informatics (LIPIcs), vol. 149, pp. 12:1–12:14. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2019). https://doi.org/10.4230/LIPIcs.ISAAC.2019.12, https://drops.dagstuhl.de/opus/volltexte/2019/11508
20. Kannan, S., Naor, M., Rudich, S.: Implicit representation of graphs. In: Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing. p. 334–343. STOC '88, Association for Computing Machinery, New York, NY, USA (1988). https://doi.org/10.1145/62212.62244, https://doi.org/10.1145/62212.62244
21. Kawarabayashi, K.i., Klein, P.N., Sommer, C.: Linear-space approximate distance oracles for planar, bounded-genus and minor-free graphs. In: Aceto, L., Henzinger, M., Sgall, J. (eds.) Automata, Languages and Programming. pp. 135–146. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
22. Liu, Y.: Topological Theory of Graphs:. De Gruyter (2017). https://doi.org/doi:10.1515/9783110479492, https://doi.org/10.1515/9783110479492
23. Peleg, D.: Informative labeling schemes for graphs. In: Nielsen, M., Rovan, B. (eds.) Mathematical Foundations of Computer Science 2000. pp. 579–588. Springer Berlin Heidelberg, Berlin, Heidelberg (2000)
24. Seufert, S., Anand, A., Bedathur, S., Weikum, G.: Ferrari: Flexible and efficient reachability range assignment for graph indexing. In: 2013 IEEE 29th International Conference on Data Engineering (ICDE). pp. 1009–1020 (2013). https://doi.org/10.1109/ICDE.2013.6544893

25. Simon, K.: An improved algorithm for transitive closure on acyclic digraphs. In: Kott, L. (ed.) Automata, Languages and Programming. pp. 376–386. Springer Berlin Heidelberg, Berlin, Heidelberg (1986)
26. Thomassen, C.: Kuratowski's theorem. Journal of Graph Theory **5**(3), 225–241 (1981). https://doi.org/https://doi.org/10.1002/jgt.3190050304, https://onlinelibrary.wiley.com/doi/abs/10.1002/jgt.3190050304
27. Thorup, M.: Compact oracles for reachability and approximate distances in planar digraphs. J. ACM **51**(6), 993–1024 (Nov 2004). https://doi.org/10.1145/1039488.1039493, https://doi.org/10.1145/1039488.1039493
28. Wang, H., He, H., Yang, J., Yu, P., Yu, J.: Dual labeling: Answering graph reachability queries in constant time. In: 22nd International Conference on Data Engineering (ICDE'06). pp. 75–75 (2006). https://doi.org/10.1109/ICDE.2006.53
29. Wei, H., Yu, J.X., Lu, C., Jin, R.: Reachability querying: an independent permutation labeling approach. The VLDB Journal **27**(1), 1–26 (Feb 2018). https://doi.org/10.1007/s00778-017-0468-3, https://doi.org/10.1007/s00778-017-0468-3
30. Yildirim, H., Chaoji, V., Zaki, M.J.: Grail: Scalable reachability index for large graphs. Proc. VLDB Endow. **3**(1–2), 276–284 (Sep 2010). https://doi.org/10.14778/1920841.1920879, https://doi.org/10.14778/1920841.1920879