Reconstructing a Bounded-Degree Directed Tree Using Path Queries

Zhaosen Wang
Google
Tokyo, Japan
zhaoz@google.com

Jean Honorio
Purdue University
West Lafayette, IN, USA
jhonor@purdue.edu

Abstract—We present a randomized algorithm for reconstructing directed rooted trees of \( n \) nodes and node degree at most \( d \), by asking at most \( O(dn \log^2 n) \) path queries. Each path query takes as input an origin node and a target node, and answers whether there is a directed path from the origin to the target. Regarding lower bounds, we show that any randomized algorithm requires at least \( \Omega(n \log n) \) queries, while any deterministic algorithm requires at least \( \Omega(dn) \) queries. Additionally, we present a \( O(dn \log^3 n) \) randomized algorithm for noisy queries, and a \( O(dn \log^2 n) \) randomized algorithm for additive queries on weighted trees.

I. INTRODUCTION

Scientists in diverse areas, such as statistics, epidemiology and economics, aim to unveil relationships within variables from collected data. This process can be seen as the task of reconstructing a graph (i.e., finding the hidden edges of a graph) by asking queries to an oracle. In this graph, each vertex is a variable, and each edge denotes the relationship between two variables. For instance, in cancer research, biologists try to discover the causal relationships between genes. By providing a specific treatment to a particular gene (origin), biologists can observe whether there is an effect in another gene (target). This effect can be either direct (if the two genes are connected with a directed edge) or indirect (if there is a directed path from the origin to the target gene.) In the above example, we can think of nature as the oracle that is answering whether there is a directed path from an origin to a target node. Arguably, the cost of asking queries is very high in several application domains. Thus, we are interested on the reconstruction of graphs that do not require the trivial \( n^2 \) queries for \( n \) nodes (i.e., one query for every possible pair of nodes.)

Prior work on reconstructing graphs has exclusively focused on undirected graphs. In [1], a \( O(dn \log n) \) algorithm was provided for recovering undirected trees of \( n \) nodes and maximum node degree \( d \), by using queries that return the path length between two given nodes. Authors of [2] provide a \( O(dn) \) algorithm for reconstructing undirected trees of \( n \) nodes and maximum node degree \( d \), and for a query that returns the distance from a given node to every other node. The results in [3], [4], [5] pertain to queries that answer whether there exists at least one edge between a given set of nodes. While [3], [4] focused on matchings and stars, the work of [5] provides a \( O(m \log n) \) algorithm for directed graphs of \( n \) nodes and \( m \) edges.

The results in [6], [7] pertain to queries that return the number of edges between a given set of nodes. A \( O(dn) \) algorithm was provided in [6] for undirected graphs of \( n \) nodes and maximum node degree \( d \), while a \( O(m \log n) \) algorithm was given in [7] for undirected graphs of \( n \) nodes and \( m \) edges. Other works have focused on the recovery of weighted undirected graphs. The work of [8] provides a \( O(dn \log n) \) algorithm for recovering weighted undirected trees of \( n \) nodes and maximum node degree \( d \), by using queries that return the sum of edge weights on the path between two given nodes. The work of [9], [10] pertains to the reconstruction of weighted undirected graphs of \( n \) nodes and \( m \) edges. A \( O(m \log n) \) algorithm was provided for a query that gives the sum of edge weights between a given set of nodes.

The closest work to ours is [11], which provides a \( O(dn \log^2 n) \) randomized algorithm for undirected trees of \( n \) nodes and maximum node degree \( d \), and for separator queries. A separator query takes three nodes \( i, j, k \) as input, and answers whether \( k \) is on the undirected path between \( i \) and \( j \). In contrast, our work pertains to directed rooted trees. Furthermore, we use a different type of query which we call path query. A path query takes an ordered pair of nodes \( i, j \) as input, and answers whether there exists a directed path from \( i \) to \( j \).

We provide a randomized algorithm for reconstructing directed rooted trees of \( n \) nodes and node degree at most \( d \), in \( O(dn \log^2 n) \) time. To the best of our knowledge, there is no simple reduction to transform our problem to the problem of [11] or any of the above mentioned literature. Our algorithm relies on the divide and conquer approach, the use of even separators [12] and sorting. Regarding lower bounds, we show that any randomized algorithm requires at least \( \Omega(n \log n) \) queries, while any deterministic algorithm requires at least \( \Omega(dn) \) queries. We also present a \( O(dn \log^3 n) \) randomized algorithm for a noisy regime, in which the bit that represents the oracle’s answer gets flipped with some probability, by an adversary, before it is revealed to the algorithm. Furthermore, we present a \( O(dn \log^2 n) \) randomized algorithm for reconstructing weighted trees by
using additive queries that return the sum of the edge weights on the directed path between two given nodes. We finish the paper by showing some negative results that provide some motivation for our assumptions. We show that any deterministic or randomized algorithm requires at least $\Omega(n^2)$ queries in order to recover more general directed acyclic graphs. We also show that any deterministic algorithm requires at least $\Omega(n^2)$ queries for recovering a family of sparse disconnected graphs, as well as a family of sparse connected graphs.

II. Preliminaries

In this section, we provide several formal definitions which will be useful for the detailed description of our algorithm. For clarity, we also provide some preliminary introduction to the main aspects of our algorithm.

Let $G = (V, E)$ be a directed acyclic graph with vertex set $V$ and edge set $E$. For clarity, when $G$ is a directed rooted tree, we use $T$ instead of $G$. In this paper, we assume that $T$ has $n$ nodes, i.e., $|V| = n$. Furthermore, we also assume that the node degree is at most $d$. (In a directed acyclic graph, the node degree is the sum of the indegree and the outdegree of the node.)

Recall that a path in $G$ from node $i$ to node $j$ (both in $V$) is a sequence of nodes $i, x_1, x_2, \ldots, x_k, j$ such that $(i, x_1), (x_1, x_2), \ldots, (x_{k-1}, x_k), (x_k, j)$ is a subset of the edge set $E$.

Our algorithm reconstructs a directed rooted tree, by using path queries. Next, we formally define path queries.

**Definition 1.** Let $G = (V, E)$ be a directed acyclic graph. A path query is a function $Q_G : V \times V \to \{0, 1\}$ such that $Q_G(i, j) = 1$ if there exists a path in $G$ from $i$ to $j$, and $Q_G(i, j) = 0$ otherwise.

Note that the above query only reveals a single bit of information, and it does not provide any information regarding the length of the path, thus making graph reconstruction a nontrivial task.

In this paper, we assume that the node set $V$ is known, while edge set $E$ is unknown. Our main problem is indeed to reconstruct $E$ by using path queries. We use $Q(i, j)$ to denote $Q_T(i, j)$ since for our problem, the directed rooted tree $T$ is fixed (but unknown).

A key step in our algorithm is the recovery of what we call multidirectional paths. A multidirectional path consists of the directed edges associated to an undirected path in the skeleton graph (i.e., the graph obtained by replacing each directed edge with an undirected edge in the original graph.) Figure 1 provides a visual illustration for intuitive understanding. Next, we formally define multidirectional paths.

**Definition 2.** Let $G = (V, E)$ be a directed acyclic graph. A multidirectional path of $G$ between nodes $i$ and $j$ is a sequence of nodes $i, x_1, x_2, \ldots, x_k, j$ such that each node in $V$ appears at most once in the sequence, and that there is an edge on either direction between each pair of adjacent nodes in the sequence. That is, either $(i, x_1) \in E$ or $(x_1, i) \in E$, either $(x_1, x_2) \in E$ or $(x_2, x_1) \in E$, ..., either $(x_{k-1}, x_k) \in E$ or $(x_k, x_{k-1}) \in E$, and either $(x_k, j) \in E$ or $(j, x_k) \in E$.

Next, we show that for directed rooted trees, a multidirectional path between any two arbitrary nodes always exists and is unique. More importantly, we show that a multidirectional path is either a directed path, or two directed paths that share the same origin (i.e., the lowest common ancestor.) Later, we leverage this property for recovering multidirectional paths.

**Lemma 1.** Let $T = (V, E)$ be a directed rooted tree. Given any two arbitrary nodes $i$ and $j$, a multidirectional path of $T$ between $i$ and $j$ always exists and is unique. Furthermore, a multidirectional path of $T$ between $i$ and $j$, is either a path from $i$ to $j$, or a path from $j$ to $i$, or two paths (one from $k$ to $i$, and one from $k$ to $j$, for some $k \in V - \{i, j\}$.) In this case, node $k$ is the lowest common ancestor of $i$ and $j$.

(See http://arxiv.org/abs/1606.05183 for detailed proofs.)

Our algorithm relies on the divide and conquer approach. In order to apply the above approach in our problem, it is important to introduce the concepts of even separators and bags. Next, we introduce even separators.
Definition 3. Let \( T = (V, E) \) be a directed rooted tree of bounded degree \( d \) and let \( n = |V| \). An even separator of \( T \) is an edge \( e \in E \) that when removed from \( T \), divides \( T \) into two subtrees \( T_1 \) and \( T_2 \), where each of the subtrees have a number of nodes between \( n/d \) and \((d - 1)n/d\).

The existence of even separators is pivotal for using divide and conquer in our problem. Corollary 2.3 in [12] shows that if a graph is a bounded-degree directed tree, then an even separator exists. For our graph reconstruction problem, once the even separator is identified, we cut the tree through the even separator. This operation splits the tree into two subtrees. We then recursively call the algorithm for both subtrees. We illustrate this in Figure 2.

While even separators exist [12], it remains to know whether they can be efficiently found. We show later that (on average) there is an even separator in the multidirectional path between two nodes chosen independently and uniformly at random (See Theorem 1.)

In what follows, we formally define bags, which are also important for the divide and conquer approach taken here.

Definition 4. Let \( T = (V, E) \) be a directed rooted tree, and \( S \) be the set of edges in a multidirectional path of \( T \). Define \( T_S = (V, E - S) \) as the subgraph of \( T \) after we remove all edges in \( S \). A bag with respect to a node \( i \) in a multidirectional path with edges \( S \), is a subset of nodes in \( V \) that contains \( i \) and all the nodes that are reachable from \( i \) in the (undirected) skeleton graph of \( T_S \).

Intuitively speaking, we can think of edges in a tree as “ropes”. If we “nail” all nodes of a multidirectional path into the “wall”, then all other nodes will “hang” from one of the nodes in the path. Nodes that hang from the same particular node belong to the same bag. We include a visual example in Figure 3.

In our algorithm, we recover exactly all the directed edges in a multidirectional path. Bags are used to count the number of nodes associated to each node in the multidirectional path (without the need to recover all the directed edges.) For each edge in the multidirectional path, one can then count the number of nodes on the two subtrees that would be generated if we were to cut the tree through the given edge. This process is used for identifying even separators.

Finally, our algorithm also performs sorting of nodes with a properly defined order relation, which is used for instance in the recovery of the directed edges in multidirectional paths.

Definition 5. Define the order relation of two nodes \( i \) and \( j \) as follows. If \( Q(i, j) = 1 \) we say that \( i \) is “less than” \( j \), and “greater than” otherwise.

III. ALGORITHM

In this section, we present our randomized algorithm and analyze its time complexity. Our algorithm is similar in spirit to [11] which applies to undirected trees and separator queries. In this paper, we focus on directed rooted trees and path queries. We remind the reader that path queries only reveal a single bit of information, and they do not provide any information regarding the length of the path. (We discuss noisy and additive extensions in Section IV.)

In what follows, we explain our divide and conquer approach in our main Algorithm 8. A high-level overview is shown in Figure 4. (See Appendix A for detailed algorithms.) First, we randomly pick two nodes and recover the sequence of nodes in the multidirectional path between those two randomly chosen nodes (Algorithm 1). Then, we divide the rest of the nodes (not in the multidirectional path) into bags defined by the multidirectional path (Algorithm 4.) Later, we count the number of nodes on each bag, and determine if there exists an even separator in the multidirectional path. We repeat the whole procedure (i.e., from choosing a new pair of random nodes) until we find an even separator. Finally, we cut the tree through the even

![Fig. 3. Four different directed rooted trees, and their bags with respect to the multidirectional path between two nodes chosen independently and uniformly at random.](image-url)
Alg. 8: Reconstruct tree $O(dn \log^2 n)$
Alg. 1: Reconstruct multidirectional path $O(n \log n)$
Alg. 4: Find bag $O(\log n)$
Alg. 2: Find lowest common ancestor $O(n \log n)$
Alg. 3: Find path from root $O(n \log n)$

Fig. 4. Main algorithm (in red), subroutines, and their time complexity.

separator, thus effectively splitting the tree into two subtrees (Algorithm 5.) We then recursively call our Algorithm 8 for both subtrees (until the input tree contains a single node.) On a more technical side, reconstructing a multidirectional path that consists of two paths that share the same origin, is more involved than reconstructing single paths. The former requires finding the lowest common ancestor of the two randomly chosen nodes (Algorithm 2) as well as finding the root of the tree (Algorithm 3.)

In our main Algorithm 8, finding out the bag of a particular node with respect to a directed path is relatively easier than with respect to a multidirectional path. Note that by Lemma 1, a multidirectional path is either a directed path, or two directed paths that share the same origin (i.e., the lowest common ancestor.) In Algorithm 8, we simplify the bag assignment task, by splitting a multidirected path (that is not a single path) into its two constituent directed paths. The algorithm first recovers the lowest common ancestor, and then breaks the multidirectional path into two paths. For instance, the multidirectional path 1, 2, 3, 4, 5 in Figure 3(d) with lowest common ancestor 3 would become two paths 3, 2, 1 and 3, 4, 5.

Next, we explain each subroutine used in our main Algorithm 8. The first key step is to recover the multidirectional path between two randomly chosen nodes. Algorithm 1 recovers the sequence of nodes in the multidirectional path between any two nodes.

The idea behind Algorithm 1 is as follows. Let $i, j \in V$ be two arbitrary nodes in the directed rooted tree $T = (V, E)$. Recall that some multidirectional paths are a single directed path. We can easily detect this case by asking whether there is a path from $i$ to $j$ (i.e., whether $Q(i, j) = 1$), or whether there is a path from $j$ to $i$ (i.e., whether $Q(j, i) = 1$). Without loss of generalization, assume there is a path in the directed rooted tree $T$ from $i$ to $j$. (We could similarly assume that there is a path from $j$ to $i$.) We can find out the set of all nodes $\{x_1, x_2, \ldots x_k\}$ on the path from $i$ to $j$, by using path queries. For all nodes $k \in V - \{i, j\}$, we ask the oracle about $Q(i, k)$ and $Q(k, j)$. Note that $k$ is on the path from $i$ to $j$, if and only if $Q(i, k) = Q(k, j) = 1$. After finding out the set of all nodes $\{x_1, x_2, \ldots x_k\}$ on the path from $i$ to $j$, it remains to sort the nodes in order to obtain the correct sequence, thus recovering the path. We sort the list of nodes $\{i, x_1, x_2, \ldots x_k, j\}$ by using the order relation given in Definition 5.

Some multidirectional paths consist of two directed paths that share the same origin $m$. In this case, Algorithm 1 first recovers the lowest common ancestor $m$ and then reconstructs two directed paths: one from $m$ to $i$, and one from $m$ to $j$, by following the approach explained before. In Algorithm 1, reconstructing a multidirectional path that consists of two paths requires finding the lowest common ancestor of the two randomly chosen nodes. Our Algorithm 2 finds the lowest common ancestor of a multidirectional path between any two arbitrary nodes.

Algorithm 2 works as follows. Let $i, j \in V$ be two arbitrary nodes of the directed rooted tree $T = (V, E)$. First, we recover the directed path from the root to $i$. We assume that the order of the nodes in the above path follow the order relation given in Definition 5. Thus, the tree root is the first element on such path. Then, we iterate through all nodes in the path, in order to find the last ancestor of $j$ in the path from the root. This last ancestor is indeed the lowest common ancestor of $i$ and $j$. In Algorithm 2, in order to find the lowest common ancestor, one has to find the root of the directed tree. Our Algorithm 3 identifies the path from the root to a given arbitrary node.

The inner workings of Algorithm 3 are as follows. Let $i \in V$ be an arbitrary node in the directed rooted tree $T = (V, E)$. For each node $j \in V - \{i\}$, we ask the oracle about $Q(j, i)$. If $Q(j, i) = 1$ then there is a path from $j$ to $i$, and therefore we add node $j$ to the list of nodes that reach $i$. In order to recover the directed path from the root to $i$, we sort the aforementioned list of nodes, by using the order relation given in Definition 5. That is, the first element on the sorted list is the tree root.

The second key step in our main Algorithm 8 is to divide the nodes (which are not in the multidirectional path) into bags defined by the multidirectional path. As argued before, if the multidirectional path consists of two directed paths, then Algorithm 8 breaks the multidirectional path into its two constituent directed paths, by using the least common ancestor. Thus, the bag assignment task needs only to consider directed paths, as we do in Algorithm 4.

Here we give an intuitive explanation of the bag assignment task in Algorithm 4. For instance, consider finding out the bag of node 10 in the directed path 1, 2, 3, 4, 5 in Figure 3(a). We see that nodes 1, 2 and 3 are all ancestors of node 10 (i.e., $Q(1, 10) = Q(2, 10) = Q(3, 10) = 1$.) We also see that nodes 4 and 5 are not ancestors of node 10 (i.e., $Q(4, 10) = Q(5, 10) = 0$.) Note that node 10 belongs to the bag of node 3. The above suggests that the task of finding out the bag of node 10 can be done by searching for a node $i$ for which $Q(i, 10) = 1$ and $Q(j, 10) = 0$ where $(i, j)$ is an edge in the path. This can be efficiently done by performing binary search.

There are two exceptions to the above rule: when queries
for all nodes in the path return 1, and when queries for all nodes in the path return 0. As an example for the first case (when all queries return 1), consider finding out the bag of node 11 in the directed path 1, 2, 3, 4, 5 in Figure 3(a). We see that nodes 1, 2, 3, 4 and 5 are all ancestors of node 11 (i.e., \(Q(1, 11) = Q(2, 11) = Q(3, 11) = Q(4, 11) = Q(5, 11) = 1\)). In this case, we assign node 11 to the bag of the last node in the path, i.e., 5. As an example for the second case (when all queries return 0), consider finding out the bag of node 10 in the directed path 3, 4, 5 in Figure 3(d). We see that neither nodes 3, 4 or 5 are ancestors of node 10 (i.e., \(Q(3, 10) = Q(4, 10) = Q(5, 10) = 0\)). In this case, we assign node 10 to the bag of the lowest common ancestor, i.e., 3. Fortunately, all the cases analyzed above are naturally handled by binary search.

The final step in our main Algorithm 8 is to cut the tree through the even separator, which splits the tree into two subtrees. Our Algorithm 5 splits a directed rooted tree into two subtrees, by cutting the original tree through any arbitrary edge.

The idea behind Algorithm 5 is as follows. Let \(e = (i, j) \in E\) be an arbitrary edge in the directed rooted tree \(T = (V, E)\). Let \(T_1\) and \(T_2\) be two subtrees that result from removing \(e\) from \(T\). Note that every node must belong to either \(T_1\) or \(T_2\). Without loss of generality, let \(i \in T_1\) and \(j \in T_2\). Since \(T\) is a directed rooted tree, \(j\) has one parent in \(T\), which is indeed \(i\). Removing \((i, j)\) from \(T\) makes \(j\) have no parents in \(T_2\). Therefore, \(j\) is the root of \(T_2\). For all nodes \(k \in V - \{j\}\), we ask the oracle about \(Q(j, k)\). If \(Q(j, k) = 1\), then \(k\) belongs to \(T_2\), otherwise \(k\) belongs to \(T_1\).

We finish the section by analyzing the time complexity of our randomized algorithm.

**Theorem 1.** Algorithm 8 takes \(O(dn \log^2 n)\) expected time, in order to reconstruct a directed rooted tree of \(n\) nodes and node degree at most \(d\). Furthermore, for a fixed probability of error \(\delta \in (0, 1)\), Algorithm 8 takes at most \(O(\frac{1}{3} dn \log^2 n)\) time, with probability at least \(1 - \delta\).

(See Appendix B for the detailed proof.)

**IV. LOWER BOUND AND EXTENSIONS**

In this section, we study lower bounds for reconstructing directed rooted trees from path queries. We also extend our original algorithm for the case of noisy queries, as well as the reconstruction of weighted trees from additive queries. Finally, we provide negative results for directed acyclic graphs that provide some motivation for our assumptions.

**A. Lower Bounds**

Here, we present lower bounds for reconstructing directed rooted trees from path queries.

**Theorem 2.** In order to reconstruct a directed rooted tree of \(n\) nodes and node degree at most \(d\), any randomized algorithm requires at least \(\Omega((1 - \delta)n \log n)\) queries, otherwise it would fail with probability at least \(\delta\).

**Theorem 3.** In order to reconstruct a directed rooted tree of \(n\) nodes and node degree at most \(d\), any deterministic algorithm requires at least \(\Omega(n \log n)\) queries.

Given the above results and Theorem 1, our algorithm is only a factor of \(O(d \log n)\) from the lower bound for any randomized algorithm, and a factor of \(O(\log^2 n)\) from the lower bound for any deterministic algorithm. Both of these factors are small when compared to the time complexity of our algorithm, which is \(O(dn \log^2 n)\).

**B. Noisy Queries**

Here, we analyze a noisy regime, in which the bit that represents the oracle’s answer gets flipped with some probability, by an adversary, before it is revealed to the algorithm. Next, we formally define noisy queries.

**Definition 6.** Let \(G = (V, E)\) be a directed acyclic graph, and let \(Q_G\) be a path query. A noisy path query with noise parameter \(\varepsilon \in (0, 1/2)\) is a function \(Q_G : V \times V \rightarrow \{0, 1\}\) such that \(Q_G(i, j) = Q_G(i, j)\) with probability \(1 - \varepsilon\), and \(\hat{Q}_G(i, j) = 1 - Q_G(i, j)\) with probability \(\varepsilon\).

In order to make use of our original algorithm, we proceed with the following strategy. Algorithm 6 works as follows. For each node pair, we perform majority voting on \(m\) noisy path queries. If \(m\) is large enough, noise will be removed with high probability. (See Appendix A for details.)

The above opens up a question on the number of queries \(m\) per node pair, that are sufficient to guarantee graph recovery success. A second question is whether \(m\) depends on the number of nodes \(n\) and maximum node degree \(d\), thus affecting the time complexity of our randomized algorithm. The following theorem answers both questions.

**Theorem 4.** For a fixed probability of error \(\delta \in (0, 1)\) and noise parameter \(\varepsilon \in (0, 1/2)\), Algorithm 6 takes at most \(O(\frac{1}{3} n \log^2 n (\log d + \log n + \log \frac{1}{\delta}))\) time, with probability at least \(1 - \delta\), in order to reconstruct a directed rooted tree of \(n\) nodes and node degree at most \(d\), provided that the number of queries per node pair fulfills \(m \in \Theta(\frac{1}{172 - \pi^2} (\log d + \log n + \log \frac{1}{\delta}))\).

In practice we do not need to know the exact value of the noise parameter \(\varepsilon\). A lower bound of \(\varepsilon\) suffices to define the number of queries \(m\) per node pair.

**C. Additive Queries on Weighted Trees**

Here, we focus on the reconstruction of weighted directed rooted trees by using additive queries. An additive path query returns the sum of the edge weights on the directed path between two given nodes, if such a path exists, or zero otherwise. Next, we formally define additive path queries.

**Definition 7.** Let \(T = (V, E, W)\) be a weighted directed rooted tree, with positive weights for each edge, i.e., \(w_{i,j} > 0\) for all \((i, j) \in E\), and \(w_{i,j} = 0\) for all \((i, j) \notin E\). An additive path query is a function \(\hat{Q}_T : V \times V \rightarrow [0, +\infty)\) such that if there exists a path in \(T\) from \(i\) to \(j\) then \(\hat{Q}_T(i, j)\)
A trivial algorithm for reconstructing directed acyclic graphs of length greater than 1. Due to this path, we have that
\[ Q(i, j) = 0 \quad \text{otherwise.} \]

Note that the above query reveals much more information compared to the path query in Definition 1 which only reveals a single bit of information. In Algorithm 7, our strategy is to convert the additive query problem into our original problem for recovering the edge set. Afterwards, we recover the edge weights by calling the additive queries for each edge. (See Appendix A for details.)

The time complexity of the above algorithm is as follows.

**Theorem 5.** Algorithm 7 takes \( O(dn \log^2 n) \) expected time, in order to reconstruct a weighted directed rooted tree of \( n \) nodes and node degree at most \( d \). Furthermore, for a fixed probability of error \( \delta \in (0, 1) \), Algorithm 7 takes at most \( O\left(\frac{1}{\delta} d n \log^2 n\right) \) time, with probability at least \( 1 - \delta \).

**D. Negative Results for Directed Acyclic Graphs**

As argued before, the cost of asking queries is very high in several application domains. Thus, we are interested on the reconstruction of graphs that do not require the trivial \( n^2 \) queries for \( n \) nodes (i.e., one query for every possible pair of nodes). A natural question is whether more general directed acyclic graphs could be recovered efficiently by asking less than \( \Omega(n^2) \) queries. Here, we provide a negative answer for the above.

First note that some directed acyclic graphs are non-identifiable by using path queries. For instance, consider the two graphs shown in Figure 5. In both graphs, we have that \( Q(1, 2) = Q(2, 3) = Q(1, 3) = 1 \). Thus, by using path queries, it is impossible to discern whether the edge \((1, 3)\) exists or not. Next, we formalize the above intuition.

**Definition 8.** Let \( G = (V, E) \) be a directed acyclic graph. We say that an edge \((i, j) \in E\) is transitive if there exists a directed path from \( i \) to \( j \) of length greater than 1.

In Figure 5, edge \((1, 3)\) is transitive, since there is a directed path \(1, 2, 3\) (i.e., a directed path of length 2.) Transitive edges are not possible to be recovered by using path queries for the following reason. Let \( i \) and \( j \) be two fixed nodes. Assume there is a directed path from \( i \) to \( j \) of length greater than 1. Due to this path, we have that \( Q(i, j) = 1 \), regardless of whether \((i, j) \in E\) or \((i, j) \notin E\). A trivial algorithm for reconstructing directed acyclic graphs of \( n \) nodes, without transitive edges, is to ask \( n^2 \) queries (i.e., one query for every possible pair of nodes.) Next, we show that this trivial algorithm is indeed optimal.

**Theorem 6.** In order to reconstruct a directed acyclic graph of \( n \) nodes, without transitive edges, any deterministic algorithm requires at least \( \Omega(n^2) \) queries. Furthermore, any randomized algorithm requires at least \( \Omega((1 - \delta)^2) \) queries, otherwise it would fail with probability at least \( \delta \).

Recall that our algorithm pertains to directed rooted trees with a maximum node degree. These graphs are not only sparse but also weakly connected (i.e., their undirected skeleton graphs are connected.) One could ask whether connectedness is a necessary condition, and whether sparsity makes graph reconstruction easier. Next, we show that an algorithm requires \( \Omega(n^2) \) queries for recovering a family of sparse disconnected graphs, as well as a family of sparse connected graphs.

**Theorem 7.** In order to reconstruct a sparse disconnected directed acyclic graph of \( n \) nodes, any deterministic algorithm requires at least \( \Omega(n^2) \) queries.

**Theorem 8.** In order to reconstruct a sparse connected directed acyclic graph of \( n \) nodes, any deterministic algorithm requires at least \( \Omega(n^2) \) queries.

V. CONCLUDING REMARKS

There are several ways of extending this research. The analysis of the reconstruction of other families of graphs in \( O(n \log n) \) time would be of great interest. Given our results for directed rooted trees, it would be interesting to analyze other families of sparse connected graphs, such as graphs with bounded tree-width as well as graphs with bounded arboricity.

**REFERENCES**

[1] J. Hein, “An optimal algorithm to reconstruct trees from additive distance data,” Bulletin of Mathematical Biology, vol. 51, no. 5, pp. 597–603, 1989.
[2] Z. Beelerova, F. Eberhard, T. Erlebach, A. Hall, M. Hoffmann, M. Mihaluk, and L. Ram, “Network discovery and verification,” IEEE Journal on Selected Areas in Communications, vol. 24, no. 12, pp. 2168–2181, 2006.
[3] N. Alon, R. Beigel, S. Kasif, S. Rudich, and B. Sudakov, “Learning a hidden matching,” SIAM Journal on Computing, vol. 33, no. 2, pp. 487–501, 2004.
[4] N. Alon and V. Asodi, “Learning a hidden subgraph,” International Colloquium on Automata, Languages and Programming, Lecture Notes in Computer Science, vol. 3142, pp. 110–121, 2004.
[5] D. Angluin and J. Chen, “Learning a hidden graph using \( O(\log n) \) queries per edge,” Conference on Learning Theory, pp. 210–223, 2004.
[6] V. Grebinski and G. Kucherov, “Optimal reconstruction of graphs under the additive model,” Algorithmica, vol. 28, no. 1, pp. 104–124, 2000.
[7] L. Reyzin and N. Srivastava, “Learning and verifying graphs using queries with a focus on edge counting,” Algorithmic Learning Theory, Lecture Notes in Computer Science, vol. 4754, pp. 285–297, 2007.
[8] J. Culberson and P. Rudnicki, “A fast algorithm for reconstructing trees from distance matrices,” Information Processing Letters, vol. 30, no. 4, pp. 215–220, 1989.
[9] S.-S. Choi, “Polynomial time optimal query algorithms for finding graphs with arbitrary real weights,” Conference on Learning Theory, pp. 1–22, 2013.
[10] H. Mazzawi, “Optimally reconstructing weighted graphs using queries,” ACM-SIAM Symposium on Discrete Algorithms, pp. 608–615, 2010.
[11] M. Jagadish and A. Sen, “Learning a bounded-degree tree using separator queries,” Algorithmic Learning Theory, Lecture Notes in Computer Science, vol. 8139, pp. 188–202, 2013.
[12] F. Chang, “Separator theorems and their applications,” Paths, Flows, and VLSI-Layout, pp. 17–34, 1990.
[13] M. Akra and L. Bazzi, “On the solution of linear recurrence equations,” Computational Optimization and Applications, vol. 10, no. 2, pp. 195–210, 1998.
[14] T. Cover and J. Thomas, Elements of Information Theory, 2nd ed. John Wiley & Sons, 2006.
Algorithm 1 Reconstruct multidirectional path

1: **Input**: vertex set \( V \) of the directed rooted tree \( T \), two nodes \( i, j \in V \).
2: multidirpath ← \([]\).
3: if \( Q(i, j) = 1 \) then
4: for each node \( k \in V - \{i, j\} \) where \( Q(i, k) = 1 \) and \( Q(k, j) = 1 \)
5: multidirpath ← append(multidirpath, \( k \))
6: end for
7: Sort multidirpath with the order relation given in Definition 5
8: multidirpath ← append(i, multidirpath, \( j \))
9: lowestancestoridx ← 1
10: else if \( Q(j, i) = 1 \) then
11: for each node \( k \in V - \{i, j\} \) where \( Q(j, k) = 1 \) and \( Q(k, i) = 1 \)
12: multidirpath ← append(multidirpath, \( k \))
13: end for
14: Sort multidirpath with the order relation given in Definition 5
15: multidirpath ← append(i, multidirpath, \( j \))
16: lowestancestoridx ← 1
17: else
18: \( m \leftarrow \text{Find lowest common ancestor}(V, i, j) \)
19: pathleft ← \([\] \), pathright ← \([\] \)
20: for each node \( k \in V - \{m, i\} \) where \( Q(m, k) = 1 \) and \( Q(k, i) = 1 \)
21: pathleft ← append(pathleft, \( k \))
22: end for
23: Sort pathleft with the order relation given in Definition 5
24: for each node \( k \in V - \{m, j\} \) where \( Q(m, k) = 1 \) and \( Q(k, j) = 1 \)
25: pathright ← append(pathright, \( k \))
26: end for
27: Sort pathright with the order relation given in Definition 5
28: multidirpath ← append(i, reverse(pathleft), \( m \), pathright, \( j \))
29: lowestancestoridx ← 2[length(pathleft)]
30: end if
31: **Output**: multidirpath, lowestancestoridx

Algorithm 2 Find lowest common ancestor

1: **Input**: vertex set \( V \) of the directed rooted tree \( T \), two nodes \( i, j \in V \).
2: pathfromroot ← **Find path from root**(\( V, i \))
3: for each node \( k \) in pathfromroot do
4: if \( Q(k, j) = 1 \)
5: lowestcommonancestor ← \( k \)
6: break
7: end if
8: end for
9: **Output**: lowestcommonancestor

Algorithm 3 identifies the path from the root to a given arbitrary node.

Algorithm 4 Find bag

1: **Input**: path \( x_1, x_2, \ldots, x_k \), node \( i \in V \) where \( V \) is the vertex set of the directed rooted tree \( T \)
2: \( l \leftarrow 1 \), \( r \leftarrow k \)
3: while \( l < r \) do
4: \( m \leftarrow (l + r)/2 \)
5: if \( Q(x_m, i) = 1 \) then
6: \( l \leftarrow m \)
7: else
8: \( r \leftarrow m \)
9: end if
10: end while
11: **Output**: bag index \( l \in \{1, \ldots, k\} \)

Algorithm 5 splits a directed rooted tree into two subtrees, by cutting the original tree through an arbitrary edge.

Algorithm 6 Reconstruct weighted tree from additive path queries. We use \( \tilde{Q}^{(k)}(i, j) \) to denote the \( k \)-th call to the query \( \tilde{Q}(i, j) \), since for our problem, the directed rooted tree \( T \) is fixed (but unknown).

Algorithm 7 Reconstruct weighted tree from additive queries

1: **Input**: vertex set \( V \) of the weighted directed rooted tree \( T \), vertex \( i \in V \)
2: Define the path query \( Q \) as follows. Let \( Q(i, j) = 1 \) if \( \sum_{k=1}^{m} \tilde{Q}^{(k)}(i, j) > m/2 \), and \( Q(i, j) = 0 \) otherwise.
3: \( E \leftarrow \text{Reconstruct tree}(V) \)
4: **Output**: edge set \( E \)

Algorithm 7 reconstructs a weighted directed rooted tree by using additive path queries. We use \( \hat{Q}(i, j) \) to denote \( \hat{Q}(i, j) \) since for our problem, the weighted directed rooted tree \( T \) is fixed (but unknown).

Algorithm 8 reconstructs a directed rooted tree by using path queries.
Algorithm 8 Reconstruct tree
1: Input: vertex set $V$ of the directed rooted tree $T$
2: if $|V|=1$ then
3: $E \leftarrow \emptyset$
4: else
5: while true do
6: Pick $i, j \in V$ independently and uniformly at random
7: $(\text{multidirpath}, \text{lowestancestoridx}) \leftarrow \text{Reconstruct multidirectional path}(V, i, j)$
8: Assume multidirpath $= [x_1, \ldots, x_p]$
9: pathleft $\leftarrow \text{reverse(multidirpath[1 to lowestancestoridx])}$
10: pathright $\leftarrow \text{multidirpath[lowestancestoridx to } P]\}$
11: bagsize $\leftarrow \{1, \ldots, p\}$, an array of size $p$
12: for each node $k$ that is not on multidirpath do
13: bagidxleft $\leftarrow \text{Find bag}(\text{pathleft}, k)$
14: bagidxright $\leftarrow \text{Find bag}(\text{pathright}, k)$
15: if bagidxleft $= \text{lowestancestoridx}$ then
16: bagidx $\leftarrow \text{bagidxright} + \text{length(pathleft)} - 1$
17: else
18: bagidx $\leftarrow \text{bagidxleft}$
19: end if
20: bagsize[bagidx] $\leftarrow$ bagsize[bagidx] + 1
21: end for
22: leftsize $\leftarrow 0$
23: evenseparator $\leftarrow $ None
24: for $r = 1, \ldots, P - 1$ do
25: leftsize $\leftarrow$ leftsize + bagsize[r]
26: if leftsize $\in [n/d, (d - 1)n/d]$ then
27: evenseparator $\leftarrow (x_r, x_{r+1})$
28: break
29: end if
30: end for
31: if evenseparator $\neq $ None then
32: break
33: end if
34: end while
35: $(V_1, V_2) \leftarrow \text{Split tree}(V, \text{evenseparator})$
36: $E_1 \leftarrow \text{Reconstruct tree}(V_1)$
37: $E_2 \leftarrow \text{Reconstruct tree}(V_2)$
38: $E \leftarrow E_1 \cup E_2 \cup \{\text{evenseparator}\}$
39: end if
40: Output: edge set $E$

APPENDIX
PROOF OF THEOREM 1

Proof. Let a “round” be a repetition of the “while” loop at Line 5 of Algorithm 8. We show that there are $O(d)$ rounds in expectation, and that each round takes $O(n \log n)$ time. We then finish the proof by the application of the master theorem.

First, we analyze the expected number of rounds for finding an even separator in the multidirectional path between two nodes chosen independently and uniformly at random. Recall that by Lemma 1 for directed rooted trees, a multidirectional path between any two arbitrary nodes always exists and is unique. Thus, it remains to analyze the probability that two randomly chosen nodes lie on a different subtree of the directed rooted tree $T$. Since both nodes are selected independently and uniformly at random from the set of $n$ nodes, then $q^2$ is the probability that both nodes fall in the first subtree. Similarly, $(1 - q)^2$ is the probability that both nodes fall in the second subtree. The probability $p$ that the two nodes lie on a different subtree is

$$p \geq \min_{q \in [1/d, (d-1)/d]} \left(1 - q^2 - (1 - q)^2\right) = 1 - (1/d)^2 - ((d-1)/d)^2 = 2(d-1)/d^2 \in \Omega(1/d).$$

Therefore, the expected number of rounds until we successfully find two nodes lying on a different subtree is $E[r] = \sum_{r=1}^{\infty} r(1-p)^{r-1} = 1/p$ which is $O(d)$. We now show that each round takes $O(n \log n)$ time. First, we derive a series of conclusions regarding the subtrees:

- Algorithm 3 takes $O(n \log n)$ time, since the most time-consuming step is sorting in Line 8.
- Algorithm 2 takes $O(n \log n)$ time, since the most time-consuming step is the call to Algorithm 3 in Line 2, which takes $O(n \log n)$ time.
- Algorithm 1 takes $O(n \log n)$ time, since the most time-consuming steps are sorting in Lines 7, 14, 23 and 27, and the call to Algorithm 2 in Line 18, which takes $O(n \log n)$ time.
- Algorithm 4 takes $O(\log n)$ time, since it performs binary search.
- Algorithm 5 takes $O(n)$ time, since the “for” loop in Line 3 iterates for at most $n$ times.

Recall that each round calls Algorithms 1, 4 and 5. It can be observed then, that the most time-consuming step on each round is the call to Algorithm 1, which takes $O(n \log n)$ time.

Thus, so far we know that the time complexity of the “while” loop at Line 5 is $O(dn \log n)$. To finalize the proof, note that Algorithm 8 exits the “while” loop at Line 5 when it finds an even separator. Recall from Definition 3 that the even separator splits the tree into two subtrees, where each of the subtrees have a number of nodes between $n/d$ and $(d - 1)n/d$. The total running time for Algorithm 8 is given by the recursive formula

$$C(n) = C(n/d) + C((d-1)n/d) + O(dn \log n).$$

For clarity, we rewrite $C(n)$ in terms of the master theorem in [13]. That is, $C(n) = \alpha_1 C(\beta_1 n) + \alpha_2 C(\beta_2 n) + \gamma(n)$, for $\alpha_1 = \alpha_2 = 1$, $\beta_1 = 1/d$, $\beta_2 = (d - 1)/d$ and $\gamma(n) \in O(dn \log n)$. By invoking the master theorem in [13], we have that $C(n) \in O(n^{s \log s})$, where $s$ is the value for which $\alpha_1 \beta_1^s + \alpha_2 \beta_2^s = 1$. In our case $s = 1$ and thus $C(n) \in O(dn \log^2 n)$.

For the remainder of the proof, we use $C$ to denote $C(n)$ since $n$ is a constant. Note that $C$ is a non-negative random variable with expectation $E[C] \in O(dn \log^2 n)$. By Markov’s inequality, we have that $P[C > a] \leq E[C]/a$. By letting $a = E[C]/\delta$, we have $P[C > \delta] \leq \delta$. Therefore $P[C \leq \delta] \geq 1 - \delta$, and we prove our claim that $P[C \leq \delta] \geq 1 - \delta$. \qed