Relating Two Property Testing Models for Bounded Degree Directed Graphs

Artur Czumaj[*]
Dept. Computer Science
Centre for Discrete Mathematics
and its Applications (DIMAP)
University of Warwick
United Kingdom
A.Czumaj@warwick.ac.uk

Pan Peng[†]
Dept. Computer Science
TU Dortmund, Germany &
State Key Lab of Computer
Science, Institute of Software,
Chinese Academy of Sciences
pan.peng@tu-dortmund.de

Christian Sohler
Dept. Computer Science
TU Dortmund, Germany
christian.sohler@tu-dortmund.de

ABSTRACT

We study property testing algorithms in directed graphs (digraphs) with maximum indegree and maximum outdegree upper bounded by d. For directed graphs with bounded degree, there are two different models in property testing introduced by Bender and Ron (2002). In the *bidirectional model*, one can access both incoming and outgoing edges while in the *unidirectional model* one can only access outgoing edges. In our paper we provide a new relation between the two models: we prove that if a property can be tested with *constant query complexity* in the bidirectional model, then it can be tested with *sublinear query complexity* in the unidirectional model.

A corollary of this result is that in the unidirectional model (the model allowing only queries to the outgoing neighbors), every property in hyperfinite digraphs is testable with sublinear query complexity.

Categories and Subject Descriptors

F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems

General Terms

Theory

Keywords

graph property testing; sampling based algorithms; directed graph algorithms

1. INTRODUCTION

A fundamental task in the study of large networks such as the web graph, social networks, etc., is to analyze their structural properties. For example, we may want to know if a network is well-connected or has many copies (instances) of some specific sub-structures. Since many modern networks are massive and of quickly growing size, the problem of performing the structural analysis efficiently has been becoming increasingly important and even linear time algorithms are often too slow for this task. In such a scenario, one of the most viable approaches is (random) sampling. However, sampling involves many challenges. First of all, there are different sampling methods and some may be better or worse for the task at hand. Secondly, there is the question how to interpret the sample, i.e., what can we learn about the structure of the graph from our random sample. These questions require a systematic treatment, which is done in the area of property testing, a formal framework to study sampling algorithms for the analysis of structural properties of large graphs. *Property testing* is a relaxation of classical decision problems that aims to distinguish between objects having a predetermined property (e.g., graphs being well-connected) and objects being far from any object having the property (e.g., graphs being poorly-connected). The notion of being "far" is problem dependent; one typically assumes that the algorithm rejects objects that are ε-far from having Π, where an object is ε-far from property Π if one has to modify more than an ε fraction of its representation to obtain an object with property Π. For example, if the input graph $G = (V, E)$ is represented by adjacency lists, then G is called ε-far from property Π (say, planarity) if one has to modify more than $\varepsilon|E|$ edges in G to obtain a graph with property Π.

The notion of property testing was first formulated by Rubinfeld and Sudan [26], as it arises naturally in the context of program verification and learning theory. Goldreich et al. [14] initiated the study of property testing for graphs and combinatorial objects, and in the recent years we have seen numerous property testing algorithms to test various graph properties (see, e.g., [13, 25] and the references therein).

The area of graph property testing has been extensively studied in the last two decades, with many beautiful results combining graph algorithms, extremal graph theory, theory of random graphs, and complexity theory. While there has been comprehensive research of graph property testing in the context of undirected graphs (see, e.g., surveys [13, 25]

[*] Supported in part by the Centre for Discrete Mathematics and its Applications (DIMAP) and by EPSRC grant No. EP/J021814/1.

[†] The second and third author acknowledge the support by ERC grant No. 307696.

STOC'16, June 19–21, 2016, Cambridge, MA, USA
ACM. 978-1-4503-4132-5/16/06...$15.00
http://dx.doi.org/10.1145/2897518.2897575

and references in Section 1.2 on related work), surprisingly little successful efforts have been devoted to the study of *directed graphs.* The main goal of this work is to advance our understanding of testing properties of *bounded degree directed graphs* (digraphs).

There are two most natural models of accessing (bounded degree) *digraphs*, as introduced by Bender and Ron [5]: the *bidirectional model* that allows to query outgoing and incoming edges of a vertex, and the *unidirectional model* that allows to query only the outgoing edges (see, e.g., [5, 12, 20])[1]. Depending on the context, either model can be more appropriate in any given situation. These two models are quite different though. The former model resembles the model of undirected graphs (see, e.g., [12]), is at least as strong as the second model, and has some non-trivial testers with constant query complexity, cf. [5, 23, 27]. The latter model is on one hand more natural in applications and for graph exploration, and on the other hand it is algorithmically more challenging, and achieving even sublinear query complexity for testing is highly nontrivial. A representative scenario where one can see a big difference between these two models is when one processes web graphs, in which each vertex u corresponds to a webpage and a directed edge (u, v) corresponds to a hyperlink from the webpage corresponding to u to the webpage corresponding to v. Such graphs are relatively easy to be explored in the model allowing to query outgoing and incoming edges of a vertex, but in many applications the queries for incoming edges are not allowed or are too expensive. The unidirectional model is more natural, but it is significantly more complicated to analyze the input network in this model, since, for example, it is impossible to quickly learn about any incoming edge of any vertex.

The main goal of this paper is to *study the relation between these two models* and *provide a generic transformation that converts testers in the bidirectional model, to testers with sublinear query complexity in the unidirectional model that allows only queries to the outgoing edges.*

1.1 Description of New Results

A *d-bounded digraph* is a digraph with both maximum outdegree and indegree upper bounded by d. A *digraph property* Π is defined by a set of digraphs closed under digraph isomorphism (renaming of vertices). A d-bounded digraph G is called *ε-far from satisfying property* Π if one has to modify more than εdn edges of G to obtain a d-bounded degree directed graph G' satisfying Π. Given a digraph property Π, our goal is to design a randomized algorithm, called *tester*, that for every $\varepsilon > 0$ can distinguish, with probability at least $\frac{2}{3}$, between digraphs satisfying Π and digraphs that are ε-far from satisfying Π, while making as few queries as possible. Testers may have one-sided or two-sided error. A tester has *one-sided error* if it accepts every digraph satisfying Π, and so it can err (with probability at most $\frac{1}{3}$) only for digraphs that are ε-far from satisfying Π. A tester that can err (with probability at most $\frac{1}{3}$) for digraphs satisfying Π and also for digraphs that are ε-far from satisfying Π, is said to be with *two-sided error.*

Definition 1. Let $\Pi = (\Pi_n)_{n\in\mathbb{N}}$ be a d-bounded digraph property $\Pi = (\Pi_n)_{n\in\mathbb{N}}$, where Π_n is a property of d-bounded digraphs with n vertices. We call Π to be *q-query testable with error probability δ* (for some function $q = q(n, \varepsilon, d)$), if for every n, ε and d there exists a tester that makes $q = q(n, \varepsilon, d)$ queries and with probability at least $1 - \delta$, accepts any n-vertex d-bounded digraph G satisfying Π and rejects any n-vertex d-bounded digraph G that is ε-far from satisfying Π [2]. If $\delta = \frac{1}{3}$, then we simply say that Π *is q-query testable*, or that Π *can be tested with query complexity q.*

Following the main line of research in graph property testing, we will assume that both ε and d are constant, even though we will parameterize our analysis with respect to these two parameters. With this in mind, we will say that a tester has *constant query complexity* if its query complexity is a function of ε and d only, and is independent of n. Furthermore, throughout the paper we use the notation $\Omega_{\varepsilon,d}()$ to describe a function in the Big-Omega notation assuming that ε and d are constant. And so, as the result, the query complexity $n^{1-\Omega_{\varepsilon,d}(1)}$ will imply that the complexity is *sublinear* in n, assuming that ε and d are constant.

Main result.

We provide a generic transformation between the two main models of access to digraphs.

Theorem 1.1. *A graph property that can be tested with two-sided error and query complexity $O_{\varepsilon,d}(1)$ in the bounded degree digraph bidirectional model can be tested with $n^{1-\Omega_{\varepsilon,d}(1)}$ queries and two-sided error in the bounded degree digraph unidirectional model.*

We remark that the two-sided error feature of the tester in the unidirectional model is necessary to ensure our result. Indeed, for example, it is known that strong connectivity can be tested with one-sided error with constant number of queries in the bidirectional model [5], whereas there is no one-sided error tester for strong connectivity with query complexity $o(n)$ in the unidirectional model [12, 20].

Let us also note that the bound $n^{1-\Omega_{\varepsilon,d}(1)}$ for the query complexity cannot be improved much and we would not be surprised if this bound was tight. Indeed, it is known that one can test 3-star-freeness with a constant number queries in the model allowing outgoing and incoming neighbor queries, while any tester with two-sided error in the model that only allows outgoing neighbor queries requires $\Omega(n^{2/3})$ queries [20], which directly implies the same lower bound for our result: there are graph properties with constant query complexity in the bounded degree digraph bidirectional model that require $\Omega(n^{2/3})$ queries in the bounded degree digraph unidirectional model.

Testing hyperfinite properties.

As a corollary of our main theorem, we show that a large class of digraph properties, called *hyperfinite digraph properties*, can be tested with sublinear query complexity in the unidirectional model that allows only outgoing neighbor queries.

Definition 2. An undirected graph G is called *(ε, k)-hyperfinite* if it can be partitioned into connected components of size at most k each, after removing at most εn edges. For

[1] There is nothing special to restrict to outgoing neighbor queries, and the model that permits *only incoming neighbor queries* can be considered similarly.

[2] Notice that the tester may depend on n, ε and d, similarly to previous works, see, e.g., [15, 16, 21]

a function $\rho : \mathbb{R}^+ \to \mathbb{R}^+$, G is called *ρ-hyperfinite* if G is $(\varepsilon, \rho(\varepsilon))$-hyperfinite for any $\varepsilon > 0$. A class of undirected graphs is called *ρ-hyperfinite* if all the graphs in the class are ρ-hyperfinite. A class of undirected graphs is called *hyperfinite* if there exists a function $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ such that it is ρ-hyperfinite.

These definitions can be applied to directed graphs by a natural extension. For this purpose let us define the underlying undirected graph of a directed graph to be the graph obtained by replacing all directed edges by undirected edges (if edges are present in both direction we replace them by a single edge).

Definition 3. A directed graph G is called a *(ε, k)-hyperfinite digraph* if the underlying undirected graph of G is (ε, k)-hyperfinite. For a function $\rho : \mathbb{R}^+ \to \mathbb{R}^+$, G is called *ρ-hyperfinite* if G is $(\varepsilon, \rho(\varepsilon))$-hyperfinite for any $\varepsilon > 0$. A class of digraphs is called *ρ-hyperfinite* if all the digraphs in the class are ρ-hyperfinite. A class of digraphs is called *hyperfinite* if there exists a function $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ such that it is ρ-hyperfinite.

Note that the class of hyperfinite graphs contains many natural classes of graphs, e.g., all planar bounded degree graphs, all bounded degree graphs defined by a finite collection of forbidden minors, etc (cf. [19, 21]).

Newman and Sohler [21] showed that every graph property of a hyperfinite undirected graph is testable with constant query complexity in the bounded degree (undirected) graph model, where the testing algorithm in the bounded degree graph model is allowed to query the neighbors of any given vertex. This together with Theorem 1.1 implies the following theorem.

THEOREM 1.2. *Every graph property of a hyperfinite digraph is testable with query complexity $O_{\varepsilon,d}(1)$ in the bounded degree bidirectional model and with query complexity $n^{1-\Omega_{\varepsilon,d}(1)}$ in the bounded degree unidirectional model.*

PROOF SKETCH. We only have to prove that every graph property of a hyperfinite digraph is testable with query complexity $O_{\varepsilon,d}(1)$ in the bidirectional model. Then the theorem follows from Theorem 1.1.

It is known that every graph property of hyperfinite graphs is testable with query complexity $O_{\varepsilon,d}(1)$ [21] and the result here follows by an easy modification of the proofs, which we only sketch here.

The proof in [21] uses the concept of local partitioning oracles introduced by [19]. A local partitioning oracle $\mathcal{O}_G$ of G provides access to a partition of $V(G)$ such that for any query about v, it computes the partition class of $V(G)$ containing v by making only a constant number of queries to G; the partition depends only on the graph G and random bits of the oracle (see [19, 22] for details).

We can extend the construction of local graph partitioning oracle for hyperfinite graphs to hyperfinite digraphs by applying the same construction as in [19], with a constant number of queries to incoming and outgoing neighbors. This follows from the fact that the underlying undirected graph is hyperfinite and only vertices within constant distance to the query vertex v will be explored by a local partitioning oracle.

This implies that the oracle provides access to a digraph G' that has connected components of constant size and is, say, $\varepsilon/4$-close to G, i.e., it differs from an isomorphic copy of G in at most $\varepsilon dn/4$ edges. As observed in [21] using a constant number of samples one can approximate the distribution of these small components and in this way obtain a constant size description of a digraph G'' that is $\varepsilon/2$-close to G (see also [22] for a related approach in a learning setting). We can then accept, if G'' is $\varepsilon/2$-close to the tested property and reject otherwise. □

Furthermore, other characterization results from [21] can be obtained similarly for hyperfinite digraphs. Let us call a graph (or digraph) property *hyperfinite* if it contains only hyperfinite graphs (or digraphs, respectively). We have the following theorem.

THEOREM 1.3. *Every hyperfinite digraph property is testable with constant query complexity in the bounded degree bidirectional model and query complexity $n^{1-\Omega_{\varepsilon,d}(1)}$ in the bounded degree unidirectional model.*

PROOF. Let Π be the considered property of digraphs. We first define an undirected property Π' that contains every graph G for which at least one orientation of its edges is in Π. Clearly, Π' is hyperfinite and so it follows from [21] that we can test with constant query time whether G is in Π'. If the tester rejects, then we can reject. Otherwise the graph is hyperfinite and we can apply Theorem 1.2 to conclude the claim. □

1.1.1 Our Techniques

We prove our main theorem, Theorem 1.1, by first showing that in the bidirectional model, any tester for some property Π_n on n-vertex digraphs with constant query complexity can be transformed into a so-called *canonical tester*. A canonical tester first samples a small number of vertices (the roots), then for each root explores the subgraph induced by all vertices at distance at most k for some appropriate chosen constant k, where the distance between two vertices u, v in a digraph is defined to be the distance between u, v in the underlying undirected graph. Such a rooted subgraph is also called k-disc. Based on the union of the explored k-discs, the tester makes a deterministic decision whether to accept the input digraph or not. This reduction closely follows a similar reduction in undirected graphs by Goldreich and Ron [15].

Given that the in- and outdegree is bounded by a constant, any canonical tester having constant query complexity can observe only the union of a constant number of k-discs for some constant k. Furthermore, for sufficiently large n these discs are disjoint with high probability. Thus, the decision made by any canonical tester can be fully characterized by a set $\mathcal{F}_n$ of a constant number of small (rooted) graphs, each of which is a union of a constant number of k-discs, such that the tester accepts if and only if the explored subgraph is not isomorphic to any digraph in $\mathcal{F}_n$. Since, with probability at least $\frac{2}{3}$, the tester distinguishes digraphs satisfying a property Π_n from digraphs that are ε-far from satisfying Π_n, we conclude that the frequency of occurrences of sets of graphs from $\mathcal{F}_n$ in graphs satisfying Π_n is small and it is large in digraphs that are ε-far from satisfying Π_n. Therefore, to simulate the tester in the model that allows only queries to outgoing neighbors, it suffices to approximate the frequencies of all k-discs in G and then calculate for any graph from $\mathcal{F}_n$ its probability to occur as a sample of the canonical tester. The technical difficulty is that there is no way to identify if a k-disc around a vertex v in the graph G' induced by the sampled edges contains all edges of the corresponding k-disc around v in G. As an illustration consider a star with three incoming edges as a forbidden 1-disc. If the input graph has many stars with, say, d incoming edges but no stars with exactly three incoming edges, it is likely that a sufficiently big sample contains many stars with exactly three incoming edges, although there is no such k-disc in G. This may lead to the wrong decision to assume that the graph does not have the considered property. One can resolve this by estimating the number of occurrences of stars with more incoming edges (this is, in fact, used in [20]). In our case, the situation is more complicated since we do not only want to estimate the indegree but the frequency of k-discs. Our approach to resolve this issue is to define a partial ordering among k-discs that allows us to compute an estimate knowing the estimates of all prior k-discs in this order. This way, we obtain an approximation of the frequencies of all k-discs. Finally, we conclude our analysis by noting that the edge sampling can be easily simulated in our model.

Testers in digraphs with arbitrary maximum degree.

It may be tempting to claim that actually the approach presented in this paper could be applied to digraphs with arbitrary maximum degrees. Indeed, assuming that one has appropriate access to the input digraph, one can come up with a useful characterization of digraph properties testable in the bidirectional model with query complexity $O_\varepsilon(1)$ in the framework of canonical testers (similar to that in Lemma 3.2). However, this result on its own does not suffice: it is impossible to estimate the frequencies of appropriated discs. In fact, we can prove that there are some properties in general digraphs that are constant-query testable in the bidirectional model (assuming that one has appropriate access to the digraph, for example, by querying the i^{th} neighbor of any vertex or by querying vertex degrees; one could also allow only access to a random incident edge) and that require an almost linear number of queries in the unidirectional model (that may allow outgoing neighbor queries and outdegree queries).

CLAIM 1.4. *In general digraphs, there exists a digraph property that is constant-query testable in the bidirectional model and requires $n^{1-O(\sqrt{\log\log n/\log n})}$ queries in the unidirectional model.*

PROOF. Let $\delta > 0$. We call a digraph (with arbitrary degrees) δ*-incoming* if the fraction of vertices with non-zero indegree is at least δ. A digraph $G = (V, E)$ is ε-far from being δ-incoming if one has to add at least $\varepsilon|E|$ edges to make it δ-incoming.

First, let us note that in the bidirectional model it is easy to test the property of being δ-incoming with $O(1/\delta)$ queries (for any $\delta > \varepsilon$ and $|E| = \Omega(n)$), just by sampling at random $O(1/\delta)$ vertices and then estimating the fraction of sampled vertices with non-zero incoming edges.

Next, we consider the problem in the unidirectional model. We will reduce the problem of approximating the number of distinct elements in a sequence of length n (called DISTINCT-ELEMENTS, cf. [24]) to the problem of testing if a graph is 2ε-incoming for any constant $\varepsilon > 0$. The problem of DISTINCT-ELEMENTS [24] is to establish the number of queries (for balls colors) from a sequence of n balls, each of a single color, to approximate the number of distinct colors. Raskhodnikova et al. [24] proved that one needs to query the colors of at least $n^{1-O(\sqrt{\log\log n/\log n})}$ balls to distinguish inputs with at least $\frac{n}{11}$ colors from inputs with at most $\frac{n}{40}$ colors.

Given an instance of DISTINCT-ELEMENTS of n balls, we define a digraph G on $2n$ vertices $\{1, \ldots, 2n\}$ as follows. The first n vertices correspond to n balls and the remaining n vertices correspond to n different colors (that contain all ball-colors and possibly redundant colors). If the i^{th} ball has color corresponding to the j^{th} color, then add edge $\langle i, n+j\rangle$. Note that the number of vertices with non-zero indegree is exactly the number of distinct colors.

Let $\mathcal{A}$ be any algorithm in the unidirectional model that tests if a digraph is 2ε-incoming or is ε-far from being 2ε-incoming (for any constant ε with $0 < \varepsilon \le \frac{1}{40}$). By invoking $\mathcal{A}$ on the digraph G that corresponds to the DISTINCT-ELEMENTS instance, we can distinguish inputs with at least $2\varepsilon n$ colors from inputs with less than εn colors. Furthermore, note that G can be constructed on-the-fly as follows: when $\mathcal{A}$ queries the (only) outgoing neighbor of some vertex i for some $i \le n$, if the color c_i of the i^{th} ball has already been assigned to the j^{th} vertex for some $n+1 \le j \le 2n$, then add edge $\langle i, j\rangle$ to G; otherwise, assign c_i to an arbitrary vertex j with $n+1 \le j \le 2n$ that has no assigned color yet, and then add edge $\langle i, j\rangle$ to G. The lower bound on the query complexity of DISTINCT-ELEMENTS implies the same lower bound for the query complexity of $\mathcal{A}$ in the unidirectional model. This completes the proof. □

Notice that the construction above uses only digraphs with the bounded outdegrees. A similar construction can be also shown for digraphs with bounded indegrees (e.g., of testing directed star $K_{1,n-1}$).

1.2 Related Work

Given the importance of directed graphs in a variety of applications, it is rather surprising that we have seen only a limited amount of research on *property testing in directed graphs*. The study of directed graphs in the context of property testing has been initiated by Bender and Ron [5], who introduced the main computational models and demonstrated that there is a large complexity gap in the testing of the two main models of bounded degree digraphs: The bidirectional model that allows to query vertices for their incoming and outgoing edges, and the unidirectional model that permits queries only for the outgoing edges. Bender and Ron [5] showed that while strong connectivity can be tested in the former model with $\widetilde{O}(1/\varepsilon)$ queries, in the latter model one needs $\Omega(\sqrt{n})$ queries, even when allowing *two-sided error*. Goldreich [12, Appendix A.3] and independently, Hellweg and Sohler [20], noted that the arguments in the lower bound of $\Omega(\sqrt{n})$ can be extended to obtain a lower bound of $\Omega(n)$ in the *one-sided error* model. Further, for two-sided error, Goldreich [12, Appendix A.3], and independently Hellweg and Sohler [20], designed testers for strong connectivity with $n^{1-\Omega_{\varepsilon,d}(1)}$ queries to the outgoing edges. We remark that while the techniques used in [20] for testing strong connectivity share similarity to our analysis for approximating the histogram of small subgraphs (cf. Section 4), in that birthday-problem arguments for estimating collisions are used, in [20] only a degree distribution is approximated, and here we consider the distribution of all rooted subgraphs of constant radius. In [20], the authors also presented a tester for subgraph-freeness with $O(n^{1-\frac{1}{k}})$ queries to outgoing neighbors, where k is the number of connected components in the subgraph that have no incoming edges. In particular, if a digraph H is strongly connected, then one can test if a given digraph is H-free with a constant number of queries, with one-sided error. Hellweg and Sohler [20] gave also a lower bound of $\Omega(n^{2/3})$ for testing 3-star freeness. The model allowing querying vertices for the incoming and outgoing edges has been also investigated for some specific graph properties, and besides strong connectivity [5], there has been a study of testing Eulerianity [23], k-edge connectivity [23, 27], k-vertex connectivity [23]; all these testers heavily rely on the access to both incoming and outgoing edges and achieve constant query complexity. It is known though that in this model acyclicity can not be tested with $o(n^{1/3})$ queries [5].

We also note that directed graphs have been also studied in the model of dense directed graphs (not considered in our paper), where Alon and Shapira [2] investigated the property of testing subgraph freeness.

Significantly more efforts have been put in the study of *undirected* graphs (see, e.g., [1, 3, 4, 6, 7, 8, 9, 10, 13, 14, 18, 21, 25]). In a work most closely related to ours, there is a related study on characterization of property testers in bounded degree undirected graphs, in which the testing algorithm is allowed to perform neighbor queries. Newman and Sohler [21] proved that every hyperfinite property is testable with a constant number of queries, which improves upon the result by Benjamini et al. [6], who showed that every minor-closed property of bounded degree graphs is testable with a constant number of queries. As we mentioned above (cf. Theorem 1.2), the corresponding digraph properties can also be tested in bounded degree digraph models. Goldreich and Ron gave a full characterization of constant-query *proximity-oblivious* testers with one-sided error in the bounded degree graph model [15] and in particular, they provide canonical testers for constant-query testable properties in bounded degree graphs with one sided error. Further discussions on proximity-oblivious testers have been given in [12, 17].

Goldreich and Ron studied sample-based testers that only sample elements independently from some distribution over the tested object (e.g., graphs, functions) [16]. For example, they proved that any property that can be tested by a proximity-oblivious tester with constant detection probability that makes q uniformly distributed queries can be tested by a sample-based testers of sample complexity $O(n^{1-1/q})$. This result and its proof do not generalize to ours and in particular, in the bounded degree graph model, the class of graph properties that can be tested by proximity-oblivious testers with constant query complexity is rather restricted when compared to the class of constant-query testable properties by standard testers (see [15]). Fischer et al. [11] proved recently that non-adaptive testers making a constant number of queries, over a fixed alphabet, can be tested by sample-based testers with query complexity sublinear in n; the results from [11] do not apply to our setting, since one needs a non-constant alphabet size in order to represent bounded degree digraphs.

2. PRELIMINARIES

For a directed graph $G = (V, E)$, $V = \{1, \ldots, n\}$ and a vertex $v \in V$, let $d_{\text{out}}(v)$ denote the outdegree of v and $d_{\text{in}}(v)$ denote the indegree of v. We assume that there is a degree bound d such that $d_{\text{out}}(v) \leq d$ and $d_{\text{in}}(v) \leq d$ and that d is known to the algorithm. We consider two models describing query access to the input digraph G, in one model (called the *bidirectional model*) we assume the access through an oracle that allows queries to *outgoing and incoming edges/neighbors* and in the other model (called the *unidirectional model*), we assume the access through an oracle that allows queries only to *outgoing edges/neighbors*. For an outgoing neighbor query (v, i, out), the oracle returns the endpoint of the ith outgoing edge of v if $i \leq d_{\text{out}}(v)$ and a special symbol "$\perp$" otherwise; for an incoming neighbor query (v, i, in), the oracle returns the endpoint of the ith incoming edge of v if $i \leq d_{\text{in}}(v)$ and "$\perp$" otherwise.

For a directed graph G, we will use $V(G)$ and $E(G)$ to denote its vertex set and edge set, respectively.

We define the *distance between two vertices* u, v in a digraph to be the distance between u, v in the underlying undirected graph (the length of the shortest path between u, v in the underlying undirected graph).

A (weakly) connected digraph G is called a *digraph with r roots* if it has exactly r vertices in G, say $v_1, \ldots, v_r$, marked as *roots*. In this case, we also say that G is a *digraph rooted at* $v_1, \ldots, v_r$. Given a parameter $k \geq 1$ and a d-bounded digraph $G = (V, E)$, a *k-disc rooted at a vertex $v \in V$*, denoted by $\text{disc}_k(v)$, is the subgraph (with v marked as root) of G induced by the vertices that are distance at most k from v. The *k-discs of G* are all possible $\text{disc}_k(v)$, for any $v \in V(G)$. Finally, let $s_{d,k}$ denote the maximum number of vertices in any k-disc, $s_{d,k} \leq 1 + 2d + \cdots + (2d)^k$.

3. CANONICAL TESTERS IN THE BIDIRECTIONAL MODEL

In this section, we study testers in the model that allows both incoming and outgoing neighbor queries. A central feature of this model is that one can quickly explore the k-disc rooted at any vertex v. Indeed, since we consider d-bounded digraphs, every vertex is incident to at most $2d$ edges (at most d incoming edges and at most d outgoing edges), and since we have direct access to these edges, a simple run of breadth first search (BFS) of depth $k + 1$ starting at v on the underlying undirected graph will explore all edges of $\text{disc}_k(v)$ using at most $(2d)^{k+1}$ queries. Using this observation, we can prove that in this model, every tester with constant query complexity can be transformed into a *canonical tester* that works as follows:

- uniformly samples a constant number of vertices, then
- explores the union of bounded discs rooted at the sampled vertices, and then
- makes deterministic decision whether to accept, based on an isomorphic copy of the explored subgraph.

The proof of the correctness of this transformation (Lemmas 3.1 and 3.2) follows similar arguments used earlier for transforming testers in dense graph into nonadaptive testers (cf. [18, Section 4]) and transforming testers with one-sided error in bounded degree graph model into canonical testers (cf. [15, Claim 5.5.2]). For the sake of completeness, we prove these claims below.

We call two digraphs H_1, H_2 with multiple distinguished vertices (roots) *isomorphic* to each other if and only if there is a roots-preserving isomorphism function[3] from the vertex set of H_1 to the vertex set of H_2.

We begin with the following lemma.

Lemma 3.1. ***(Canonical tester)*** *Let $\Pi = (\Pi_n)_{n\in\mathbb{N}}$ be a digraph property that can be tested in the bidirectional model with query complexity $q = q(\varepsilon, d)$ and error probability at most $\frac{1}{6}$. Then for every ε and d, there exists an infinite sequence $\mathcal{F}' = (\mathcal{F}'_n)_{n\in\mathbb{N}}$ such that for any n,*

- *$\mathcal{F}'_n$ is a set of digraphs with q roots that are the union of q (not necessarily disjoint) q-discs, and*
- *the property Π_n on n-vertex digraphs can be tested (with the error probability at most $\frac{1}{3}$) as follows:*
  - *first uniformly sample q vertices (without replacement),*
  - *then explore the q-discs rooted at the sample vertices, and*
  - *then accept the input graph if and only if the explored subgraph (with q sampled vertices as its roots) is not isomorphic to any $F \in \mathcal{F}'_n$.*

Proof. Our proof follows closely the arguments from the proof of Claim 5.5.2 from [15].

Let $\mathcal{T}$ be a tester for Π_n on n-vertex digraphs with error probability at most $\frac{1}{6}$. We will first convert $\mathcal{T}$ into a tester $\mathcal{T}_1$ that samples q vertices (marked as roots) and then returns the output on the basis of the subgraph explored by the union of q-discs rooted at sampled vertices. The value output by the tester $\mathcal{T}_1$ may depend on the labels (or identities) of explored vertices and the random coins for sampling vertices. Then we convert $\mathcal{T}_1$ into a tester $\mathcal{T}_2$ whose output depends only on the edges and non-edges in the explored subgraph, the ordering of all explored vertices and possibly its own coins; it is independent of the coins used to sample vertices. Next we convert $\mathcal{T}_2$ into a tester $\mathcal{T}_3$ whose output is independent of the ordering of all explored vertices. Finally, we convert $\mathcal{T}_3$ into a tester $\mathcal{T}_4$ that returns the output deterministically according to the unlabeled version of the explored subgraph and its roots.

$\mathcal{T}_1$: Let $\mathcal{T}_1$ be the tester that first uniformly samples (without replacement) a set S of q vertices and then explores the subgraph that is a union of q-discs of all vertices in S. We mark all vertices in S as roots. We then use $\mathcal{T}_1$ to emulate the execution of $\mathcal{T}$ in the following way. Given incoming and outgoing neighbor query access to an n-vertex digraph $G = (V, E)$, the tester $\mathcal{T}_1$ will select on-the-fly and a random, uniformly distributed permutation $\pi : V \to V$ and provide oracle access to the permuted digraph $\pi(G) = (V, \pi(E))$, where $\pi(E) := \{\langle\pi(u), \pi(v)\rangle : \langle u, v\rangle \in E\}$. Initially, all vertices in S are considered as unused. In the emulation, when $\mathcal{T}$ makes a query $(v, i, \text{out/in})$, if v has not appeared in any prior query or answer, then the tester $\mathcal{T}_1$ allocates to v an unused vertex u in the sample set S, and we let $\pi(v) = u$ and u will then be considered as used; otherwise $\mathcal{T}_1$ uses the allocation $\pi(v)$ determined before. Now if w is the answer of the query $(\pi(v), i, \text{out/in})$ and w has been selected as the image of some vertex in the permutation π before, then $\mathcal{T}_1$ returns $\pi^{-1}(w)$; otherwise, $\mathcal{T}_1$ returns a random unused value (vertex label) x and we let $\pi(x) = w$. Now if $w \in S$, then w will be considered used. The returned values will then be fed to $\mathcal{T}$. The tester $\mathcal{T}_1$ makes the same decision as the final decision of $\mathcal{T}$ after receiving all the necessary query answers.

Note that all the answers to the queries $(v, i, \text{out/in})$ used by $\mathcal{T}_1$ are vertices in one of the explored q-discs. In addition, the execution of $\mathcal{T}_1$ on G corresponds to an execution of $\mathcal{T}$ on a random isomorphic copy of G, since π is a random permutation of vertices of G. By the fact that Π_n is a digraph property invariant under digraph isomorphism, we can use similar arguments as the proof of Lemma 4.1 in [18] to show that $\mathcal{T}_1$ is a tester for Π_n that preserves error probability. Note that $\mathcal{T}_1$ makes decisions based on the explored subgraph, and may depend on the labels and the internal randomness used by the algorithm.

$\mathcal{T}_2$: Let p_S be the probability that the vertex set S is sampled, that is, $p_S = 1/\binom{n}{q}$. Let α_S be the sequence of queries and answers (of the form $(v, i, \text{out/in}) = u$) when exploring the q-discs of all vertices in S. Note that each α_S corresponds to a sequence α of queries and answers in which vertices are relabeled according to some canonical order. (For example, relabel the first queried vertex in α_S as 1, and for each query and answer $(v, i, \text{out/in}) = u$ in α_S such that v has been relabeled as j, if u has not been relabeled yet, then it is relabeled as $j + 1$ and the corresponding query and answer in α is $(j, i, \text{out/in}) = j + 1$; otherwise, the corresponding query and answer in α is $(j, i, \text{out/in}) = k$ where k is the relabeling of u that has been determined before.) We let $\mathcal{T}_2$ be the tester from $\mathcal{T}_1$ that decides only according to the relabeled sequence α it gets. More precisely, for any fixed (relabeled) answer sequence α, let $q_{S,\alpha}$ be the probability that $\mathcal{T}_1$ accepts the input when having selected S and seeing a relabeled answer sequence α. For each such sequence α, let $q_\alpha := \sum_S p_S q_{S,\alpha}$. Then $\mathcal{T}_2$ is obtained from $\mathcal{T}_1$ by making $\mathcal{T}_2$ to accept with probability q_α for every α. Then, similarly to the proof of Claim 4.2 in [18], we can show that $\mathcal{T}_2$ preserves the error probability of $\mathcal{T}_1$. Note that the execution of $\mathcal{T}_2$ does not depend on the identities of explored vertices, but possibly depends on the ordering of explored vertices.

$\mathcal{T}_3$: $\mathcal{T}_3$ accepts with probability that is equal to the average of all acceptance probabilities of $\mathcal{T}_2$ that are associated with each relabelling of vertices of the q-discs. More specifically, recall that q_α is the probability that $\mathcal{T}_2$ accepts when seeing a relabeled sequence α of queries and answers. Let $g(\alpha)$ denote the set of all digraphs (with multiple roots) that are isomorphic to the digraph underlying α. For a digraph H, let q_H denote the expected value of q_α for α over the set $\{\alpha : H \in g(\alpha)\}$. Then, we let $\mathcal{T}_3$ accept with probability q_H when seeing an explored subgraph isomorphic to H. Observe that $\mathcal{T}_3$ accepts G with the same probability as the probability that $\mathcal{T}_2$ accepts a random isomorphic copy of G. This implies that $\mathcal{T}_3$ is a tester for Π_n that preserves error probability of tester $\mathcal{T}_2$. Note that the decision of $\mathcal{T}_3$ does not depend on the ordering of all the explored vertices, while it may depend on its own random coins.

[3] By that we mean that if R_1 is the set of roots of H_1 and R_2 is the set of roots of H_2, then there is a permutation π over $V(H_1)$ such that (i) π restricted to R_1 is a permutation of R_2 and (ii) $(v, u) \in E(H_1)$ if and only if $(\pi(v), \pi(u)) \in E(H_2)$.

$\mathcal{T}_4$: Let $\mathcal{T}_4$ be the tester obtained from $\mathcal{T}_3$ that accepts the input digraph if and only if the probability associated with the explored subgraph H is at least $\frac{1}{2}$. More precisely, recall that p_H is the probability that $\mathcal{T}_3$ accepts when seeing H as the explored subgraph that is a union of q-discs rooted at q vertices. Then $\mathcal{T}_4$ is the tester that accepts the input digraph if and only if $p_{H'} \geq \frac{1}{2}$ when seeing an explored subgraph isomorphic to H'. Similarly to the proof of Lemma 4.4 in [18], one can show that $\mathcal{T}_4$ is a tester for Π_n with error probability $\frac{1}{3}$.

Finally, we note that the decision of the tester $\mathcal{T}_4$ is deterministic once it sees the explored subgraph spanned by the union of q-discs rooted at sampled vertices. The lemma then follows by defining $\mathcal{F}'_n$ to be the set of digraphs that is a union of q q-discs on which the tester rejects. □

The tester described in Lemma 3.1 will be called a *canonical tester.*

Our next lemma shows that with high probability, the canonical tester can be made to make decisions according to a set of small graphs each of which is a union of q' *disjoint* q'-discs, which will facilitate our analysis in that it will be sufficient to consider the distribution of q'-discs of the digraph (see Section 4).

LEMMA 3.2. *Let $\Pi = (\Pi_n)_{n\in\mathbb{N}}$ be a digraph property that can be tested with query complexity $q = q(\varepsilon, d)$ in the bidirectional model. Then there exists a universal constant $c > 0$ such that for every ε and d, there is an integer n_0, and an infinite sequence $\mathcal{F} = (\mathcal{F}_n)_{n\geq n_0}$, such that for any $n \geq n_0$, $\mathcal{F}_n$ is a set of digraphs, each being a union of cq disjoint (cq)-discs, and for any n-vertex digraph G,*

- *if G satisfies Π_n, then with probability* at most $\frac{5}{12}$ *the union of $(c \cdot q)$-discs rooted at $c \cdot q$ uniformly sampled (without replacement) vertices span a digraph isomorphic to one of the members in $\mathcal{F}_n$;*
- *if G is ε-far from satisfying Π_n, then with probability* at least $\frac{7}{12}$ *the $(c \cdot q)$-discs rooted at $c \cdot q$ uniformly sampled (without replacement) vertices span a digraph isomorphic to one of the members in $\mathcal{F}_n$.*

PROOF. We first amplify the error probability of the tester of $\Pi = (\Pi_n)_{n\in\mathbb{N}}$ to be at most $\frac{1}{6}$. This can be done by repeating the tester c times, for certain constant c, and then taking the majority. Since the resulting tester for Π has query complexity $q' := c \cdot q$ and error probability at most $\frac{1}{6}$, we can apply the proof of Lemma 3.1 to find the corresponding set $\mathcal{F}'_n$ such that each element of $\mathcal{F}'_n$ is a subgraph of a union of q' q'-discs. Let $\mathcal{F}_n$ be the subset of $\mathcal{F}'_n$ that contains all directed subgraphs with exactly q' disjoint weakly connected components; note that each element of $\mathcal{F}_n$ is a union of q' *disjoint* q'-discs.

If G satisfies Π_n, then by Lemma 3.1, with probability at most $\frac{1}{3}$, if we randomly (without replacement) select q' vertices from G then the q' q'-discs rooted at the sampled vertices span a digraph isomorphic to one of the members in $\mathcal{F}'_n$. Since $\mathcal{F}_n \subseteq \mathcal{F}'_n$, these q'-discs span a digraph isomorphic to some digraph from $\mathcal{F}_n$ with probability at most $\frac{1}{3} < \frac{5}{12}$.

Similarly, by Lemma 3.1, if G is ε-far from satisfying Π_n, the subgraph spanned by q' q'-discs rooted at randomly (without replacement) sampled vertices of G is isomorphic to one of the members in $\mathcal{F}'_n$ with probability at least $\frac{2}{3}$. Now let $n_0 := 12 \cdot s_{d,q'} \cdot (q')^2$ and note that for any $n \geq n_0$, with probability at most $\frac{s_{d,q'}\cdot(q')^2}{n} \leq \frac{1}{12}$, at least one pair of all q' sampled q'-discs intersects. This further implies that the probability that the corresponding subgraph contains less than q' weakly connected components is at most $\frac{1}{12}$. Therefore, the subgraph spanned by all the sampled q' vertices and their q'-discs is isomorphic to one of the members in $\mathcal{F}_n$ with probability at least $1 - \frac{1}{3} - \frac{1}{12} \geq \frac{7}{12}$. □

4. APPROXIMATING HISTOGRAM OF K-DISCS BY SAMPLING RANDOM EDGES

In this section, we provide foundations to our use of the framework developed in Lemma 3.2 and consider the problem of approximating the frequencies of occurrences (histogram) of all of k-discs (rooted at any possible vertex) in an input digraph, in the model allowing only *sampling of random edges.* We will present an algorithm that approximates, with arbitrarily small additive error, the histogram of k-discs in a digraph. Our algorithm invokes the process of sampling edges of the input digraph independently at random with some given probability p, and uses the outcome of the sampling to estimate the histogram. We can implement such a process by identifying the j^{th} outgoing edge of vertex i with the number $d(i-1)+j$, for $1 \leq i \leq n$ and $1 \leq j \leq d$. Then we sample each element from the set $\{1, \dots, nd\}$ with probability p and query the edges corresponding to the sampled elements. This enables us to implement the algorithm in the unidirectional model with the same query complexity.

Before we present details of our algorithm, let us first introduce some basic notation. The key challenge in our analysis is to avoid over-counting, and therefore in our analysis we will have to study the dependencies between subgraphs of k-discs. We say two rooted digraphs Γ_1, Γ_2 are of the same *isomorphic type* (which we will denote $\Gamma_1 \simeq \Gamma_2$) if they are isomorphic, that is, if there is a bijection $f : V(\Gamma_1) \to V(\Gamma_2)$ that is *root-preserving* (if u is the root of Γ_1 then $f(u)$ is the root of Γ_2) and such that $(u, v) \in E(\Gamma_1)$ if and only if $(f(u), f(v)) \in E(\Gamma_2)$. Let $\mathcal{H}_{d,k}$ denote the set of all isomorphic types Γ of d-bounded k-discs. That is, each $\Gamma \in \mathcal{H}_{d,k}$ is a rooted digraph with exactly one root, maximum indegree and outdegree at most d, and all vertices of Γ are within distance at most k to the root. Let $m_{d,k} = |\mathcal{H}_{d,k}|$ denote the number of all possible such types.

4.1 Ordering of K-Disc Isomorphic Types

Let us define a binary relation $\succeq$ on $\mathcal{H}_{d,k}$ such that $\Gamma' \succeq \Gamma$ if and only if Γ is a rooted subgraph of a rooted digraph isomorphic to Γ'. That is, $\Gamma' \succeq \Gamma$ if and only if there is an injection $f : V(\Gamma) \to V(\Gamma')$ that is root-preserving and such that if $(u, v) \in E(\Gamma)$ then $(f(u), f(v)) \in E(\Gamma')$. Note that $\succeq$ defines a partial order on the elements of $\mathcal{H}_{d,k}$. Let us fix any linear extension of this partial order and reorder the elements of $\mathcal{H}_{d,k}$ according to the linear extension, so that $\mathcal{H}_{d,k} = \{\Gamma_1, \Gamma_2, \dots, \Gamma_{m_{d,k}}\}$, where $\Gamma_i \succeq \Gamma_j$ implies $i \leq j$.

Let Γ and Γ' be any two k-disc types with $\Gamma' \succeq \Gamma$. We will estimate the probability of obtaining a copy of a digraph isomorphic to Γ by sampling edges of Γ', where each edge of Γ' is sampled randomly with probability p and independently from all other edges, for some $p \in [0, 1]$. Let $\lambda(\Gamma|p, \Gamma')$ be the probability that the sampled edges span a subgraph that is isomorphic to Γ; $\lambda(\Gamma|p, \Gamma) = p^{|E(\Gamma)|}$. We have the following simple lemma.

Lemma 4.1. *If $\Gamma' \succeq \Gamma$ then $\lambda(\Gamma|p,\Gamma') \leq \binom{|E(\Gamma')|}{|E(\Gamma)|} \cdot p^{|E(\Gamma)|}$. Furthermore, given Γ and Γ', $\lambda(\Gamma|p,\Gamma')$ can be calculated exactly.*

Proof. Let $\mathcal{G}_{\Gamma'}$ be the set of all subgraphs of Γ' on the vertex set $V(\Gamma')$. For every $\Gamma^* \in \mathcal{G}_{\Gamma'}$, the probability that the process of sampling independently every edge from Γ' at random with probability p will obtain digraph Γ^* is equal to $p^{|E(\Gamma^*)|} \cdot (1-p)^{|E(\Gamma')|-|E(\Gamma^*)|} \leq p^{|E(\Gamma^*)|}$.

Let $\mathtt{isom}_{\langle\Gamma,\Gamma'\rangle}$ be the number of digraphs in $\mathcal{G}_{\Gamma'}$ that are of the same isomorphic type as Γ, that is, are isomorphic to Γ and with the same root as Γ. Since every graph isomorphic to Γ has exactly $|E(\Gamma)|$ edges, we have $\mathtt{isom}_{\langle\Gamma,\Gamma'\rangle} \leq \binom{|E(\Gamma')|}{|E(\Gamma)|}$. Therefore $\lambda(\Gamma|p,\Gamma')$, which is the probability that the sampled edges span a subgraph that is isomorphic to Γ, is equal to $\mathtt{isom}_{\langle\Gamma,\Gamma'\rangle} \cdot p^{|E(\Gamma)|} \cdot (1-p)^{|E(\Gamma')|-|E(\Gamma)|} \leq \binom{|E(\Gamma')|}{|E(\Gamma)|} \cdot p^{|E(\Gamma)|}$.

Finally, we note that $\lambda(\Gamma|p,\Gamma')$ can be calculated exactly by counting $\mathtt{isom}_{\langle\Gamma,\Gamma'\rangle}$, which is the number of digraphs in $\mathcal{G}_{\Gamma'}$ that are of the same isomorphic type as Γ. □

4.2 Sampling Based Approximation for Histogram of K-Disc Types

We present a sampling based algorithm to approximate the number of rooted k-discs in the input digraph G that are isomorphic to every k-disc type $\Gamma \in \mathcal{H}_{d,k}$. That is, for every $\Gamma \in \mathcal{H}_{d,k}$, the algorithm approximates the number of vertices v in G with $\mathrm{disc}_k(v)$ being of the same isomorphic type as Γ (the histogram of k-disc types of the digraph). Our algorithm accesses the input graph only by randomly sampling its edges.

For any $1 \leq i \leq m_{d,k}$, let $\mathcal{G}(i)$ be the set of all indices j such that $\Gamma_j \succeq \Gamma_i$ and $\Gamma_j \neq \Gamma_i$, that is, $\mathcal{G}(i) = \{j : \Gamma_j \succeq \Gamma_i, \Gamma_j \neq \Gamma_i\}$. Note that our ordering of the types in $\mathcal{H}_{d,k}$ ensures that $\mathcal{G}(i) \subseteq \{1,\dots,i-1\}$.

Algorithm: $\mathtt{ApproxNumofDiscType}(G,n,d,k,\delta)$

```
1. For $i = 1$ to $m_{d,k}$ do
   (a) Sample each edge of $G$ independently at random
       with probability $p_i = \frac{\alpha_i}{(\delta^2 n)^{1/|E(\Gamma_i)|}}$, where $\alpha_i$ will
       be specified later (and is bounded by a function that
       only depends on $d,k$).
   (b) If no more than $t_i = 6m_{d,k} \cdot p_i \cdot dn$ edges are
       sampled, then
       i. compute $Y_i$ to be the number of vertices $v$ in
          the resulting sampled graph for which $\mathrm{disc}_k(v)$
          is of the same isomorphic type as $\Gamma_i$;
       ii. let $X_i = \left(Y_i - \sum_{j\in\mathcal{G}(i)} X_j \lambda(\Gamma_i|p_i,\Gamma_j)\right) p_i^{-|E(\Gamma_i)|}$
   (c) Otherwise, abort and return Fail.
2. Return $X_1,\dots,X_{m_{d,k}}$.
```

The following lemma presents key properties of the algorithm above. (Let us recall that $s_{d,k}$ denotes the maximum number of vertices in any k-disc, $s_{d,k} \leq 1 + 2d + \cdots + (2d)^k$.)

Lemma 4.2. *For any d-bounded digraph G with n vertices, parameters $\delta < 1$ and $k \geq 1$, with probability at least $\frac{2}{3}$, the algorithm ApproxNumofDiscType returns estimates $X_1,\dots,X_{m_{d,k}}$ such that $|X_i - cnt(\Gamma_i)| \leq \delta n$ for every $i \leq m_{d,k}$, where $cnt(\Gamma)$ is the number of vertices v in G for which $disc_k(v)$ is isomorphic to Γ. The algorithm samples $O_{d,k}(\delta^{-2/(d\cdot s_{d,k})} n^{1-1/(d\cdot s_{d,k})})$ edges.*

Proof. We first bound the number of edges sampled by the algorithm. (We will assume that if at any iteration i the algorithm samples more than $t_i = 6m_{d,k} \cdot p_i \cdot dn$ edges, then it will abort in Step (1c) after sampling $t_i + 1$ edges.)

Since every $\Gamma_i \in \mathcal{H}_{d,k}$ is a k-disc with maximum degree at most d, we have $|V(\Gamma_i)| \leq s_{d,k}$ and thus $|E(\Gamma_i)| \leq d \cdot s_{d,k}$. Next, we note that the algorithm performs at most $m_{d,k}$ iterations and the maximum number of edges sampled in any iteration is $1 + \max_{1\leq i\leq m_{d,k}} t_i = O(c'_{d,k} \cdot \delta^{-2/(d\cdot s_{d,k})} n^{1-1/(d\cdot s_{d,k})})$, for some $c'_{d,k}$ depending only on d,k (here we use the fact that $m_{d,k}$ and α_i are upper bounded by functions depending only on d and k). Therefore, the total number of edges sampled by the algorithm is $O_{d,k}(\delta^{-2/(d\cdot s_{d,k})} n^{1-1/(d\cdot s_{d,k})})$.

Next, we prove that with high probability, for all $i \leq m_{d,k}$ the algorithm returns a good estimate X_i of $\mathrm{cnt}(\Gamma_i)$. First note that for any i, the expected number of sampled edges in the ith iteration is $|E(G)|p_i \leq dnp_i$. Therefore, by Markov's inequality, the probability that more than $t_i = 6m_{d,k}dnp_i$ edges are sampled is at most $\frac{1}{6m_{d,k}}$. Thus, by the union bound, the probability that there is an i such that in the ith iteration more than t_i edges are sampled (i.e., the condition of Step (1c) holds) is at most $\frac{1}{6}$. From now on, we will assume that in every iteration i no more than t_i edges were sampled and charge the other case to the error probability of the tester.

Let $\kappa_{d,k} = 2^{d\cdot s_{d,k}}$, $\theta_i = (3\kappa_{d,k})^{i-m_{d,k}}$, and $\beta_i = 3^{i-m_{d,k}-2}$. Let us specify $\alpha_i = \left(\frac{8s_{d,2k}\cdot\kappa_{d,k}}{\beta_i\cdot\theta_i^2}\right)^{1/|E(\Gamma_i)|}$ in the algorithm. We first show the following claim.

Claim 4.3. *For every i, $1 \leq i \leq m_{d,k}$:*

$$\mathbf{Var}[Y_i] \leq \begin{cases} s_{d,2k} \cdot p_i^{|E(\Gamma_i)|} \cdot n & \text{if } i = 1 \ , \\ s_{d,2k} \cdot \kappa_{d,k} \cdot p_i^{|E(\Gamma_i)|} \cdot n & \text{if } i \geq 2 \ . \end{cases}$$

Proof. Let $S(\Gamma_i)$ denote the set of vertices in G whose k-discs are isomorphic to any Γ' with $\Gamma' \succeq \Gamma_i$, that is, $S(\Gamma_i) = \{v \in V(G) : \mathrm{disc}(v) \simeq \Gamma_j, j \in \mathcal{G}(i) \cup \{i\}\}$. Note that $|S(\Gamma_i)| = \sum_{j\in\mathcal{G}(i)\cup\{i\}} \mathrm{cnt}(\Gamma_j) \leq n$. Let $Z_{v,i}$ be the indicator random variable that the k-disc rooted at v in the sampled graph has type isomorphic to Γ_i. Note that $Z_{v,i} = 1$ if and only if $v \in S(\Gamma_i)$ and the sampled edges of $\mathrm{disc}(v)$ span a subgraph that is isomorphic to Γ_i, which is the event that occurs with probability $\lambda(\Gamma_i|p_i,\mathrm{disc}(v))$. Hence, recalling the definition of Y_i given in the algorithm, we have,

$$\begin{aligned}\mathbf{E}[Y_i] &= \sum_{v\in V(G)} \mathbf{E}[Z_{v,i}] = \sum_{v\in S(\Gamma_i)} \mathbf{E}[Z_{v,i}] \\ &= \sum_{j\in\mathcal{G}(i)\cup\{i\}} \mathrm{cnt}(\Gamma_j) \cdot \lambda(\Gamma_i|p_i,\Gamma_j) \ . \qquad (1)\end{aligned}$$

Note that $\lambda(\Gamma_i|p_i,\Gamma_i) = p_i^{|E(\Gamma_i)|}$ and observe that if $\mathrm{disc}(v) \simeq \Gamma_j$ for $j \in \mathcal{G}(i)$, then we obtain by Lemma 4.1

$$\begin{aligned}\mathbf{E}[Z_{v,i}] = \lambda(\Gamma_i|p_i,\Gamma_j) &\leq \binom{|E(\Gamma_j)|}{|E(\Gamma_i)|} \cdot p_i^{|E(\Gamma_i)|} \\ &\leq 2^{d\cdot s_{d,k}} \cdot p_i^{|E(\Gamma_i)|} = \kappa_{d,k} \cdot p_i^{|E(\Gamma_i)|} \ . \qquad (2)\end{aligned}$$

If we plug this inequality in the identity (1) for $\mathbf{E}[Y_i]$, then we obtain the following inequality:

$$\begin{aligned}\mathbf{E}[Y_i] &= \sum_{j\in\mathcal{G}(i)\cup\{i\}} \mathrm{cnt}(\Gamma_j)\cdot\lambda(\Gamma_i|p_i,\Gamma_j) \\ &\le \mathrm{cnt}(\Gamma_i)p_i^{|E(\Gamma_i)|} + \sum_{j\in\mathcal{G}(i)} \mathrm{cnt}(\Gamma_j)\kappa_{d,k}\cdot p_i^{|E(\Gamma_i)|} \ . \quad (3)\end{aligned}$$

Next, we will consider $\mathbf{E}[(\sum_{v\in S(\Gamma_i)} Z_{v,i})^2]$. Let $\mathrm{dt}(u,v)$ be the distance between u and v in the underlying undirected graph. Note the following,

$$\begin{aligned}\mathbf{E}[(\sum_{v\in S(\Gamma_i)} Z_{v,i})^2] &= \sum_{u\in S(\Gamma_i)}\sum_{v\in S(\Gamma_i)} \mathbf{E}[Z_{u,i}Z_{v,i}] \\ &= \sum_{u\in S(\Gamma_i)}\Big(\sum_{\substack{v\in S(\Gamma_i)\\ \mathrm{dt}(u,v)\le 2k}} \mathbf{E}[Z_{u,i}Z_{v,i}] + \sum_{\substack{v\in S(\Gamma_i)\\ \mathrm{dt}(u,v)> 2k}} \mathbf{E}[Z_{u,i}Z_{v,i}]\Big). \quad (4)\end{aligned}$$

Since for any vertex $u \in V(G)$, the number of vertices v within distance at most $2k$ is at most $s_{d,2k}$, we have that for any $u \in S(\Gamma_i)$,

$$\begin{aligned}\sum_{v\in S(\Gamma_i):\mathrm{dt}(u,v)\le 2k} \mathbf{E}[Z_{u,i}Z_{v,i}] &\le \sum_{v\in S(\Gamma_i):\mathrm{dt}(u,v)\le 2k} \mathbf{E}[Z_{u,i}] \\ &\le s_{d,2k}\cdot\mathbf{E}[Z_{u,i}] \ .\end{aligned}$$

On the other hand, if the distance between u and v is larger than $2k$, that is, if $\mathrm{dt}(u,v) > 2k$, then $Z_{u,i}$ and $Z_{v,i}$ are independent, and thus $\mathbf{E}[Z_{u,i}Z_{v,i}] = \mathbf{E}[Z_{u,i}]\cdot\mathbf{E}[Z_{v,i}]$. Hence, continuing from inequality (4), we have,

$$\begin{aligned}&\mathbf{E}[(\sum_{v\in S(\Gamma_i)} Z_{v,i})^2] \\ &\le s_{d,2k}\sum_{u\in S(\Gamma_i)}\mathbf{E}[Z_{u,i}] + \sum_{u\in S(\Gamma_i)}\mathbf{E}[Z_{u,i}]\cdot\sum_{v\in S(\Gamma_i)}\mathbf{E}[Z_{v,i}] \\ &= s_{d,2k}\cdot\mathbf{E}[Y_i] + (\mathbf{E}[Y_i])^2 \ .\end{aligned}$$

This immediately gives us the following bound for $\mathbf{Var}[Y_i]$:

$$\begin{aligned}\mathbf{Var}[Y_i] &= \mathbf{E}[Y_i^2] - (\mathbf{E}[Y_i])^2 \\ &= \mathbf{E}[(\sum_{v\in S(\Gamma_i)} Z_{v,i})^2] - (\mathbf{E}[Y_i])^2 \\ &\le s_{d,2k}\cdot\mathbf{E}[Y_i] \ .\end{aligned}$$

If we plug here identity (3), then we will obtain the following bound:

$$\mathbf{Var}[Y_i] \le s_{d,2k}(\mathrm{cnt}(\Gamma_i)p_i^{|E(\Gamma_i)|} + \sum_{j\in\mathcal{G}(i)}\mathrm{cnt}(\Gamma_j)\kappa_{d,k}p_i^{|E(\Gamma_i)|}) \ ,$$

which, after observing that $\sum_{j\in\mathcal{G}(i)\cup\{i\}}\mathrm{cnt}(\Gamma_j) \le n$, yields our bound on $\mathbf{Var}[Y_i]$, completing the proof of Claim 4.3. □

Next, we prove by induction our main claim of Lemma 4.2, that for any i, $1 \le i \le m_{d,k}$, the following holds:

$$\mathbf{Pr}[|X_i - \mathrm{cnt}(\Gamma_i)|] \ge \theta_i\delta n] \le \beta_i \ . \quad (5)$$

Before we will prove inequality (5), let us first note that (5) implies the statement of the lemma by the union bound, the fact that $\sum_{i=1}^{m_{d,k}}\beta_i = 3^{-m_{d,k}-2}\cdot\sum_{i=1}^{m_{d,k}} 3^i \le 3^{-m_{d,k}-2}\cdot\frac{3^{m_{d,k}+1}}{2} = \frac{1}{6}$, and the previous proven upper bound of $\frac{1}{6}$ on the probability that the condition of Step (1c) in any iteration is satisfied. Therefore, to complete the proof, we only have to prove inequality (5).

We begin with $i = 1$. Note that $X_1 = \frac{Y_1}{p_1^{|E(\Gamma_1)|}}$, and therefore $\mathbf{E}[X_1] = \frac{\mathbf{E}[Y_1]}{p_1^{|E(\Gamma_1)|}} = \mathrm{cnt}(\Gamma_1)$ and by Claim 4.3, $\mathbf{Var}[X_1] \le \frac{s_{d,2k}\cdot n}{p_1^{|E(\Gamma_1)|}}$. Hence, by Chebyshev's inequality,

$$\begin{aligned}\mathbf{Pr}[|X_1 - \mathrm{cnt}(\Gamma_1)| \ge \theta_1\delta n] &= \mathbf{Pr}[|X_1 - \mathbf{E}[X_1]| \ge \theta_1\delta n] \\ &\le \tfrac{\mathbf{Var}[X_1]}{\theta_1^2\delta^2n^2} \le \tfrac{s_{d,2k}\cdot n}{p_1^{|E(\Gamma_1)|}\cdot\theta_1^2\delta^2n^2} \le \beta_1 \ ,\end{aligned}$$

where the last inequality follows by our setting of p_1, θ_1, and β_1. This proves inequality (5) for $i = 1$.

Next, we will proceed by induction and prove that inequality (5) holds for $i \ge 2$, assuming that it is true for $1\ldots, i-1$. Let us first note that using Claim 4.3, by Chebyshev's inequality we have the following:

$$\begin{aligned}&\mathbf{Pr}\left[\left|\frac{Y_i}{p_i^{|E(\Gamma_i)|}} - \mathbf{E}\left[\frac{Y_i}{p_i^{|E(\Gamma_i)|}}\right]\right| \ge \frac{\theta_i\delta n}{2}\right] \\ &\le \frac{\mathbf{Var}[Y_i]}{p_i^{2|E(\Gamma_i)|}}\cdot\frac{4}{\theta_i^2\delta^2n^2} \le \frac{s_{d,2k}\cdot\kappa_{d,k}\cdot p_i^{|E(\Gamma_i)|}\cdot n}{p_i^{2|E(\Gamma_i)|}}\cdot\frac{4}{\theta_i^2\delta^2n^2} \\ &\le \frac{\beta_i}{2} \ .\end{aligned}$$

Next, by induction, we know that for each $j \le i-1$, with probability at least $1-\beta_j$ we have $|X_j - \mathbf{E}[X_j]| \le \theta_j\delta n$. Thus, with probability at least $1 - \sum_{j\in\mathcal{G}(i)}\beta_j \ge 1 - \sum_{j=1}^{i-1} 3^{j-m_{d,k}-2} \ge 1 - \frac{\beta_i}{2}$, we have

$$\begin{aligned}&\frac{|\sum_{j\in\mathcal{G}(i)}(X_j - \mathbf{E}[X_j])\cdot\lambda(\Gamma_i|p_i,\Gamma_j)|}{p_i^{|E(\Gamma_i)|}} \\ &\le \frac{\lambda(\Gamma_i|p_i,\Gamma_j)}{p_i^{|E(\Gamma_i)|}}\cdot\sum_{j\in\mathcal{G}(i)}\theta_j\delta n \le \kappa_{d,k}\cdot\sum_{j\in\mathcal{G}(i)}\theta_j\delta n \\ &\le \kappa_{d,k}\cdot\sum_{j=1}^{i-1}\theta_j\delta n \le \kappa_{d,k}\cdot\sum_{j=1}^{i-1}(3\kappa_{d,k})^{j-m_{d,k}}\delta n \\ &\le \frac{(3\kappa_{d,k})^{i-m_{d,k}}}{2}\delta n = \frac{\theta_i}{2}\cdot\delta n \ ,\end{aligned}$$

where in the second inequality we used (2), that $\lambda(\Gamma_i|p_i,\Gamma_j) \le \kappa_{d,k}\cdot p_i^{|E(\Gamma_i)|}$.

Hence, by noting that

$$\begin{aligned}&\left|\tfrac{\sum_{j\in\mathcal{G}(i)} X_j\cdot\lambda(\Gamma_i|p_i,\Gamma_j)}{p_i^{|E(\Gamma_i)|}} - \mathbf{E}\left[\tfrac{\sum_{j\in\mathcal{G}(i)} X_j\cdot\lambda(\Gamma_i|p_i,\Gamma_j)}{p_i^{|E(\Gamma_i)|}}\right]\right| \\ &= \tfrac{|\sum_{j\in\mathcal{G}(i)}(X_j - \mathbf{E}[X_j])\cdot\lambda(\Gamma_i|p_i,\Gamma_j)|}{p_i^{|E(\Gamma_i)|}} \ ,\end{aligned}$$

we obtain that with probability at least $1-\beta_i$,

$$\begin{aligned}&\left|\frac{Y_i - \sum_{j\in\mathcal{G}(i)} X_j\cdot\lambda(\Gamma_i|p_i,\Gamma_j)}{p_i^{|E(\Gamma_i)|}}\right. \\ &\qquad\left. - \mathbf{E}\left[\frac{Y_i - \sum_{j\in\mathcal{G}(i)} X_j\cdot\lambda(\Gamma_i|p_i,\Gamma_j)}{p_i^{|E(\Gamma_i)|}}\right]\right| \\ &\le \left|\frac{Y_i}{p_i^{|E(\Gamma_i)|}} - \mathbf{E}\left[\frac{Y_i}{p_i^{|E(\Gamma_i)|}}\right]\right| \\ &\qquad + \left|\frac{\sum_{j\in\mathcal{G}(i)} X_j\cdot\lambda(\Gamma_i|p_i,\Gamma_j)}{p_i^{|E(\Gamma_i)|}}\right. \\ &\qquad\left. - \mathbf{E}\left[\frac{\sum_{j\in\mathcal{G}(i)} X_j\cdot\lambda(\Gamma_i|p_i,\Gamma_j)}{p_i^{|E(\Gamma_i)|}}\right]\right|\end{aligned}$$

$$\leq \quad \frac{\theta_i \delta n}{2} + \frac{\theta_i \delta n}{2} = \theta_i \delta n \ . \tag{6}$$

Therefore, using the fact that

$$\mathrm{cnt}(\Gamma_i) = \mathbf{E}\left[\left(\left(Y_i - \sum_{j \in \mathcal{G}(i)} X_j \cdot \lambda(\Gamma_i | p_i, \Gamma_j)\right)\right)\right] \cdot p_i^{-|E(\Gamma_i)|},$$

inequality (6) yields $|X_i - \mathrm{cnt}(\Gamma_i)| \leq \theta_i \delta n$. □

5. TESTING PROPERTIES IN SUBLINEAR TIME IN THE UNIDIRECTIONAL MODEL

Now we are ready to prove our main theorem and show that using the framework developed in Lemmas 3.2 and 4.2, we can transform any tester with constant query complexity in the bounded degree digraph model that allows incoming and outgoing neighbor queries into a tester with query complexity $n^{1-\Omega_{\varepsilon,d}(1)}$ with two-sided error in the bounded degree digraph model that permits only outgoing neighbor queries.

Proof of Theorem 1.1. Let $\Pi = (\Pi_n)_{n \in \mathbb{N}}$ be a property that is testable with query complexity $q = q(\varepsilon, d)$ in the bidirectional model. Let c and $n_0 = n_0(\varepsilon, d)$ be integers as in Lemma 3.2. Let $q' = c \cdot q$ and $k = q'$. Let $n_1 := (1+k) \cdot 48 \cdot (2km_{d,k})^k$. Since for $n < \max\{n_0, n_1\}$ we can trivially test Π_n with a constant number of queries, we only consider $n \geq \max\{n_0, n_1\}$.

Let $\mathcal{F}_n$ be the set of all subgraphs (unions of k disjoint k-discs) that ensure that the tester for property Π_n accepts graphs satisfying Π_n, as guaranteed in Lemma 3.2. Note that without loss of generality, we can write each $F \in \mathcal{F}_n$ as a multiset of k-discs, that is, $F = \{\Delta_1, \ldots, \Delta_k\}$, where each Δ_i is a k-disc.

Let $m = m_{d,k}$ and $\mathcal{H}_{d,k} = \{\Gamma_i\}_{i=1}^m$ be as defined in Section 4. For convenience, for two integers N, M, we let $\binom{N}{M} = 0$ if $N < M$. We will use the following algorithm TestP to test Π_n in the model that only allows outgoing neighbor queries.

```
Algorithm: TestP($G, n, d, \varepsilon, \Pi_n$)

  1. Let $k = q'$ and $\delta = \frac{1}{48 \cdot (2km_{d,k})^k}$.
  2. Invoke ApproxNumofDiscType($G, n, d, k, \delta$).
       • If it returns Fail, then abort and return Fail.
       • Otherwise, let $X_1, \ldots, X_{m_{d,k}}$ be the returned estimates.
  3. For every $F = \{\Delta_1, \ldots, \Delta_k\} \in \mathcal{F}_n$:
       • For $1 \leq i \leq m$, let $x_j$ be the number of copies
         among $\{\Delta_j\}_{j=1}^k$ that are of the same isomorphic type as $\Gamma_i$.
       • Let $\mathfrak{estim}(F) = \frac{\Pi_{i=1}^m \binom{X_i}{x_i}}{\binom{n}{k}}$.
  4. If $\sum_{F \in \mathcal{F}_n} \mathfrak{estim}(F) < \frac{1}{2}$ then output Accept.
     Otherwise, Reject.
```

By Lemma 4.2, the query complexity of the algorithm TestP is $O(c_{d,q'} \cdot \delta^{-2/(d \cdot s_{d,q'})} \cdot n^{1-1/(d \cdot s_{d,q'})}) = n^{1-\Omega_{\varepsilon,d}(1)}$ in the unidirectional model. It remains to prove that the algorithm is indeed a property tester for Π_n.

Note that by Lemma 4.2, with probability at least $\frac{2}{3}$, the algorithm ApproxNumofDiscType returns estimates such that $|X_i - \mathrm{cnt}(\Gamma_i)| \leq \delta n$ for all $1 \leq i \leq m$. In the following, we will condition on this event and we will prove that if G satisfies Π_n, then $\sum_{F \in \mathcal{F}_n} \mathfrak{estim}(F) < \frac{1}{2}$, and if G is ε-far from satisfying Π_n, then $\sum_{F \in \mathcal{F}_n} \mathfrak{estim}(F) \geq \frac{1}{2}$. This would complete the proof.

Let us first discuss the idea behind our algorithm. Algorithm ApproxNumofDiscType(G, n, d, k, δ) ensures that each returned value X_i will be very close to $\mathrm{cnt}(\Gamma_i)$ (cf. Lemma 4.2). Therefore for every $F = \{\Delta_1, \ldots, \Delta_k\} \in \mathcal{F}_n$ and the relevant $x_1, \ldots, x_m$, we will study $\mathfrak{prob}(\Delta_1, \ldots, \Delta_k) := \frac{\Pi_{i=1}^m \binom{\mathrm{cnt}(\Gamma_i)}{x_i}}{\binom{n}{k}}$, from which we will obtain the required bounds for $\sum_{F \in \mathcal{F}_n} \mathfrak{estim}(F)$.

Observe that for any multiset $\{\Delta_1, \ldots, \Delta_k\}$, the probability that the k-discs of k vertices sampled uniformly at random without replacement span a subgraph isomorphic to the subgraph corresponding to $\{\Delta_1, \ldots, \Delta_k\}$ has the *multivariate hypergeometric distribution* with parameters $n, \mathrm{cnt}(\Gamma_1), \ldots, \mathrm{cnt}(\Gamma_m), k$. That is, if for every i, $1 \leq i \leq m$, there are exactly x_i copies in the multiset $\{\Delta_1, \ldots, \Delta_k\}$ that are of the same isomorphic type as Γ_i (note that $x_1 + \cdots + x_m = k$ for any $1 \leq i \leq m$), then the probability that the subgraph spanned by k-discs of k uniformly sampled vertices is isomorphic to $\{\Delta_1, \ldots, \Delta_k\}$ is equal to $\mathfrak{prob}(\Delta_1, \ldots, \Delta_k) = \frac{\Pi_{i=1}^m \binom{\mathrm{cnt}(\Gamma_i)}{x_i}}{\binom{n}{k}}$, where we assumed $\binom{N}{M} = 0$ for $N < M$.

To study the relation between $\mathfrak{estim}(F)$ and $\mathfrak{prob}(F)$, we begin with the following auxiliary claim.

Claim 5.1. *For any i, if $|X_i - cnt(\Gamma_i)| \leq \delta n$, it holds that $|\binom{X_i}{x_i} - \binom{cnt(\Gamma_i)}{x_i}| \leq 4\delta n^{x_i}$.*

Proof. Let us first observe that the inequality trivially holds for $x_i = 0$, and it also easily holds for $x_i = 1$: $|\binom{X_i}{x_i} - \binom{\mathrm{cnt}(\Gamma_i)}{x_i}| = |X_i - \mathrm{cnt}(\Gamma_i)| \leq \delta n \leq 4\delta n^{x_i}$. Therefore, let us assume now that $x_i \geq 2$.

Let us recall a binomial identity: $\binom{N}{M} = \sum_{K=M-1}^{N-1} \binom{K}{M-1}$, which gives for $M \leq L \leq N$ the following: $\binom{N}{M} = \binom{L}{M} + \sum_{K=L}^{N-1} \binom{K}{M-1}$. Using this identity, that $\mathrm{cnt}(\Gamma_i) \leq n$, and $x_i \geq 2$, we obtain,

$$\begin{aligned}
\binom{X_i}{x_i} &\leq \binom{\mathrm{cnt}(\Gamma_i) + \lceil \delta n \rceil}{x_i} \\
&= \binom{\mathrm{cnt}(\Gamma_i)}{x_i} + \sum_{j=1}^{\lceil \delta n \rceil} \binom{\mathrm{cnt}(\Gamma_i) + j - 1}{x_i - 1} \\
&\leq \binom{\mathrm{cnt}(\Gamma_i)}{x_i} + \lceil \delta n \rceil \cdot \binom{\mathrm{cnt}(\Gamma_i) + \lceil \delta n \rceil - 1}{x_i - 1} \\
&\leq \binom{\mathrm{cnt}(\Gamma_i)}{x_i} + 2\delta n (\mathrm{cnt}(\Gamma_i) + \delta n)^{x_i - 1} \\
&\leq \binom{\mathrm{cnt}(\Gamma_i)}{x_i} + 2\delta n ((1+\delta) n)^{x_i - 1} \\
&= \binom{\mathrm{cnt}(\Gamma_i)}{x_i} + 2\delta (1+\delta)^{x_i - 1} n^{x_i} \\
&\leq \binom{\mathrm{cnt}(\Gamma_i)}{x_i} + 4\delta n^{x_i} \ ,
\end{aligned}$$

where in the last inequality, we used the fact that $(1+\delta)^{x_i-1} \le (1+\delta)^k \le 2$.

Similarly, assuming that $\mathrm{cnt}(\Gamma_i) \ge \lceil \delta n \rceil + k$, we have $\mathrm{cnt}(\Gamma_i) \ge \lceil \delta n \rceil + x_i$, and we obtain,

$$\begin{aligned}
\binom{X_i}{x_i} &\ge \binom{\mathrm{cnt}(\Gamma_i) - \lceil \delta n \rceil}{x_i} \\
&= \binom{\mathrm{cnt}(\Gamma_i)}{x_i} - \sum_{j=1}^{\lceil \delta n \rceil} \binom{\mathrm{cnt}(\Gamma_i) - j}{x_i - 1} \\
&\ge \binom{\mathrm{cnt}(\Gamma_i)}{x_i} - \lceil \delta n \rceil \binom{\mathrm{cnt}(\Gamma_i)}{x_i - 1} \\
&\ge \binom{\mathrm{cnt}(\Gamma_i)}{x_i} - 2\delta n \binom{n}{x_i - 1} \\
&\ge \binom{\mathrm{cnt}(\Gamma_i)}{x_i} - 2\delta n \cdot n^{x_i - 1} \\
&= \binom{\mathrm{cnt}(\Gamma_i)}{x_i} - 2\delta n^{x_i} \ .
\end{aligned}$$

On the other hand, in the complementary case $\mathrm{cnt}(\Gamma_i) \le \lceil \delta n \rceil + k$, we note that $\binom{\mathrm{cnt}(\Gamma_i)}{x_i} \le \binom{\lceil \delta n \rceil + k}{x_i} \le (\lceil \delta n \rceil + k)^{x_i} \le (2\delta n)^{x_i} \le 4\delta n^{x_i}$, where the third inequality follows from the fact that $n \ge n_1 = \frac{1+k}{\delta}$. Therefore since $\binom{X_i}{x_i} \ge 0$, we have $\binom{X_i}{x_i} \ge \binom{\mathrm{cnt}(\Gamma_i)}{x_i} - 4\delta n^{x_i}$.

Now we can combine all the bounds above and obtain that for $x_2 \ge 2$, the following holds,

$$\binom{\mathrm{cnt}(\Gamma_i)}{x_i} - 4\delta n^{x_i} \le \binom{X_i}{x_i} \le \binom{\mathrm{cnt}(\Gamma_i)}{x_i} + 4\delta n^{x_i} \ ,$$

what yields the claim. □

Next, consider any $F = \{\Delta_1, \dots, \Delta_k\}$ and the corresponding frequencies $x_1, \dots, x_m$. Note that there are at most k indices i with $x_i > 0$, and that $x_1 + \cdots + x_m = k$. Let $\mathcal{I} = \{i : x_i > 0, 1 \le i \le m\}$ and thus $|\mathcal{I}| \le k$ and $\prod_{i \in \mathcal{I}} n^{x_i} = n^k$. We have the following auxiliary claim.

Claim 5.2. *For any i, conditioned on $|X_i - cnt(\Gamma_i)| \le \delta n$, the following inequalities hold:*

$$\begin{aligned}
\prod_{i \in \mathcal{I}} \left(\binom{cnt(\Gamma_i)}{x_i} + 4\delta n^{x_i} \right) &< \prod_{i \in \mathcal{I}} \binom{cnt(\Gamma_i)}{x_i} + 4\delta 2^k n^k \ , \\
\prod_{i \in \mathcal{I}} \left(\binom{cnt(\Gamma_i)}{x_i} - 4\delta n^{x_i} \right) &> \prod_{i \in \mathcal{I}} \binom{cnt(\Gamma_i)}{x_i} - 4\delta 2^k n^k \ .
\end{aligned}$$

Proof. For any $i \in \mathcal{I}$, we let $y_{i,0} = \binom{\mathrm{cnt}(\Gamma_i)}{x_i}$ and $y_{i,1} = 4\delta n^{x_i}$. Then

$$\begin{aligned}
&\prod_{i \in \mathcal{I}} \left(\binom{\mathrm{cnt}(\Gamma_i)}{x_i} + 4\delta n^{x_i} \right) \\
&= \prod_{i \in \mathcal{I}} (y_{i,0} + y_{i,1}) \\
&= \sum_{i \in \mathcal{I}, j_i \in \{0,1\}} \prod_{i \in \mathcal{I}} y_{i,j_i} \\
&= \prod_{i \in \mathcal{I}} y_{i,0} + \sum_{\substack{i \in \mathcal{I}, j_i \in \{0,1\}, \\ \text{there exists } j_i = 1}} \prod_{i \in \mathcal{I}} y_{i,j_i} \\
&= \prod_{i \in \mathcal{I}} \binom{\mathrm{cnt}(\Gamma_i)}{x_i} + \sum_{\substack{i \in \mathcal{I}, j_i \in \{0,1\}, \\ \text{there exists } j_i = 1}} \prod_{i \in \mathcal{I}} y_{i,j_i} \ .
\end{aligned}$$

Now note that for any $i \in \mathcal{I}$, $y_{i,0} = \binom{\mathrm{cnt}(\Gamma_i)}{x_i} \le n^{x_i}$. Therefore, for any sequence $\{j_i\}_{i \in \mathcal{I}}$ with at least one element equal to 1, we have the following bound $\prod_{i \in \mathcal{I}} y_{i,j_i} \le 4\delta \prod_{i \in \mathcal{I}} n^{x_i} = 4\delta n^k$. Since the total number of such indices is $2^k - 1 < 2^k$, we have

$$\begin{aligned}
&\prod_{i \in \mathcal{I}} \binom{\mathrm{cnt}(\Gamma_i)}{x_i} + \sum_{\substack{i \in \mathcal{I}, j_i \in \{0,1\}, \\ \text{there exists } j_i = 1}} \prod_{i \in \mathcal{I}} y_{i,j_i} \\
&\qquad < \prod_{i \in \mathcal{I}} \binom{\mathrm{cnt}(\Gamma_i)}{x_i} + 4\delta n^k \cdot 2^k \ ,
\end{aligned}$$

which completes the proof of the first inequality. The proof of the second inequality is analogues. □

Using Claims 5.1 and 5.2, we can prove the following relation between $\mathfrak{estim}(F)$ and $\mathfrak{prob}(F)$.

Claim 5.3. *If $|X_i - cnt(\Gamma_i)| \le \delta n$ for every i, then it holds that $|\mathfrak{estim}(F) - \mathfrak{prob}(F)| \le 4\delta(2k)^k$ for every $F \in \mathcal{F}_n$.*

Proof. Let $F = \{\Delta_1, \dots, \Delta_k\} \in \mathcal{F}_n$. By Claims 5.1 and 5.2, we have

$$\begin{aligned}
\mathfrak{estim}(\Delta_1 \dots \Delta_k) &= \frac{\prod_{i \in \mathcal{I}} \binom{X_i}{x_i}}{\binom{n}{k}} \\
&\le \frac{\prod_{i \in \mathcal{I}} \left(\binom{\mathrm{cnt}(\Gamma_i)}{x_i} + 4\delta n^{x_i} \right)}{\binom{n}{k}} \\
&< \frac{\prod_{i \in \mathcal{I}} \binom{\mathrm{cnt}(\Gamma_i)}{x_i} + 4\delta 2^k n^k}{\binom{n}{k}} \\
&\le \mathfrak{prob}(\Delta_1, \dots, \Delta_k) + 4\delta(2k)^k \ ,
\end{aligned}$$

where the last inequality follows from that $\binom{n}{k} \ge (\frac{n}{k})^k$. Similarly, by Claims 5.1 and 5.2, we have,

$$\begin{aligned}
\mathfrak{estim}(\Delta_1, \dots, \Delta_k) &\ge \frac{\prod_{i \in \mathcal{I}} \left(\binom{\mathrm{cnt}(\Gamma_i)}{x_i} - 4\delta n^{x_i} \right)}{\binom{n}{k}} \\
&\ge \frac{\prod_{i \in \mathcal{I}} \binom{\mathrm{cnt}(\Gamma_i)}{x_i} - 4\delta 2^k n^k}{\binom{n}{k}} \\
&\ge \mathfrak{prob}(\Delta_1, \dots, \Delta_k) - 4\delta(2k)^k \ .
\end{aligned}$$

□

Now, we are ready to conclude our analysis by using the properties of $\sum_{F \in \mathcal{F}_n} \mathfrak{prob}(F)$ to show that conditioned on $|X_i - \mathrm{cnt}(\Gamma_i)| \le \delta n$ for all $1 \le i \le m$, the following two claims will hold:

- if G satisfies Π_n, then $\sum_{F \in \mathcal{F}_n} \mathfrak{estim}(F) < \frac{1}{2}$, and
- if G is ε-far from satisfying Π_n, then $\sum_{F \in \mathcal{F}_n} \mathfrak{estim}(F) \ge \frac{1}{2}$.

Let G be a d-bounded digraph satisfying Π. Then, by Lemma 3.2, with probability at most $\frac{5}{12}$, the subgraph spanned by the k-discs of k vertices that are sampled uniformly at random without replacement is isomorphic to some member in $\mathcal{F}_n$, that is, $\sum_{F\in\mathcal{F}_n}\mathfrak{prob}(F)\le\frac{5}{12}$. Therefore, by Claim 5.3, we have,

$$\begin{aligned}\sum_{F\in\mathcal{F}_n}\mathfrak{estim}(F) &< \sum_{F\in\mathcal{F}_n}\mathfrak{prob}(F)+\sum_{F\in\mathcal{F}_n}4\delta(2k)^k\\ &\le \sum_{F\in\mathcal{F}_n}\mathfrak{prob}(F)+m_{d,k}^k\cdot 4\delta(2k)^k\\ &\le \frac{5}{12}+\frac{1}{12}=\frac{1}{2}\ .\end{aligned}$$

Similarly, by Lemma 3.2, if G is ε-far from satisfying Π, then with probability at least $\frac{7}{12}$, the k-discs rooted at k vertices that are sampled uniformly at random span a subgraph in $\mathcal{F}_n$. Hence, Claim 5.3 gives

$$\begin{aligned}\sum_{F\in\mathcal{F}_n}\mathfrak{estim}(F) &\ge \sum_{F\in\mathcal{F}_n}\mathfrak{prob}(F)-\sum_{F\in\mathcal{F}_n}4\delta(2k)^k\\ &\ge \sum_{F\in\mathcal{F}_n}\mathfrak{prob}(F)-m_{d,k}^k\cdot 4\delta(2k)^k\\ &\ge \frac{7}{12}-\frac{1}{12}=\frac{1}{2}\ .\end{aligned}$$

These inequalities conclude the analysis of `ApproxNumofDiscType` and the proof of Theorem 1.1. □

6. CONCLUSIONS

In this paper, we study the relationship between two property testing models for bounded degree digraphs by showing that every constant-query testable property in the bidirectional model that allows both incoming and outgoing neighbor queries can be tested in sublinear query complexity in the unidirectional model that only permits outgoing neighbor queries. The underlying transformation is performed through first characterizing by canonical testers all constant-query testable properties in the bidirectional model, and then by an analysis of approximating the histogram of d-bounded k-discs by a sampling based algorithm.

7. REFERENCES

[1] Noga Alon, Eldar Fischer, Ilan Newman, and Asaf Shapira. A combinatorial characterization of the testable graph properties: It's all about regularity. *SIAM Journal on Computing*, 39(1):143–167, 2009.

[2] Noga Alon and Asaf Shapira. Testing subgraphs in directed graphs. *Journal of Computer and System Sciences*, 69(3):354–382, 2004.

[3] Noga Alon and Asaf Shapira. A characterization of the (natural) graph properties testable with one-sided error. *SIAM Journal on Computing*, 37(6):1703–1727, 2008.

[4] Noga Alon and Asaf Shapira. Every monotone graph property is testable. *SIAM Journal on Computing*, 38(2):505–522, 2008.

[5] Michael A Bender and Dana Ron. Testing properties of directed graphs: Acyclicity and connectivity. *Random Structures & Algorithms*, 20(2):184–205, 2002.

[6] Itai Benjamini, Oded Schramm, and Asaf Shapira. Every minor-closed property of sparse graphs is testable. *Advances in Mathematics*, 223(6):2200–2218, 2010.

[7] Artur Czumaj, Oded Goldreich, Dana Ron, C. Seshadhri, Asaf Shapira, and Christian Sohler. Finding cycles and trees in sublinear time. *Random Structures & Algorithms*, 45(2):139–184, 2014.

[8] Artur Czumaj, Pan Peng, and Christian Sohler. Testing cluster structure of graphs. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC'2015)*, pages 723–732, 2015.

[9] Artur Czumaj, Asaf Shapira, and Christian Sohler. Testing hereditary properties of nonexpanding bounded-degree graphs. *SIAM Journal on Computing*, 38(6):2499–2510, 2009.

[10] Artur Czumaj and Christian Sohler. Testing expansion in bounded-degree graphs. In *Proceedings of the 48th Annual Symposium on Foundations of Computer Science (FOCS'2007)*, pages 570–578, 2007.

[11] Eldar Fischer, Oded Lachish, and Yadu Vasudev. Trading query complexity for sample-based testing and multi-testing scalability. In *Proceedings of the 56th Annual Symposium on Foundations of Computer Science (FOCS'2015)*, pages 1163–1182, 2015. Also available at arXiv:1504.00695.

[12] Oded Goldreich. Introduction to testing graph properties. In Oded Goldreich, editor, *Property Testing*, pages 105–141. Springer, 2011.

[13] Oded Goldreich, editor. *Property Testing: Current Research and Surveys*. Springer, 2011.

[14] Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, 1998.

[15] Oded Goldreich and Dana Ron. On proximity-oblivious testing. *SIAM Journal on Computing*, 40(2):534–566, 2011.

[16] Oded Goldreich and Dana Ron. On sample-based testers. In *Proceedings of the 6th Conference on Innovations in Theoretical Computer Science (ITCS'2015)*, pages 337–345, 2015.

[17] Oded Goldreich and Igor Shinkar. Two-sided error proximity oblivious testing. *Random Structures & Algorithms*, 2015. To appear. DOI: 10.1002/rsa.20582.

[18] Oded Goldreich and Luca Trevisan. Three theorems regarding testing graph properties. *Random Structures & Algorithms*, 23(1):23–57, 2003.

[19] Avinatan Hassidim, Jonathan Kelner, Huy N Nguyen, and Krzysztof Onak. Local graph partitions for approximation and testing. In *Proceedings of the 50th Annual Symposium on Foundations of Computer Science (FOCS'2009)*, pages 22–31, 2009.

[20] Frank Hellweg and Christian Sohler. Property testing in sparse directed graphs: strong connectivity and subgraph-freeness. In *Proceedings of the 20th Annual European Symposium on Algorithms (ESA'2012)*, pages 599–610. Springer, 2012. Full version appeared in arXiv:1312.0497.

[21] Ilan Newman and Christian Sohler. Every property of hyperfinite graphs is testable. *SIAM Journal on Computing*, 42(3):1095–1112, 2013.

[22] Krzysztof Onak. On the complexity of learning and testing hyperfinite graphs. 2012. Manuscript. Available at http://onak.pl/papers/.

[23] Yaron Orenstein and Dana Ron. Testing Eulerianity and connectivity in directed sparse graphs. *Theoretical Computer Science*, 412(45):6390–6408, 2011.

[24] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

[25] Dana Ron. Algorithmic and analysis techniques in property testing. *Foundations and Trends in Theoretical Computer Science*, 5(2):73–205, 2010.

[26] Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996.

[27] Yuichi Yoshida and Hiro Ito. Testing k-edge-connectivity of digraphs. *Journal of Systems Science and Complexity*, 23(1):91–101, 2010.