Explicit Upper Bounds on the Minimum Size of Planar Graphs That Satisfy a Given Distribution of k-Disks

Hendrik Fichtenberger

2. Dezember 2014

Zusammenfassung

Für einen beliebigen ungerichteten Graph G ist die k-Disk eines Knotens u als der gewurzelte Subgraph definiert, der von einer bei u gestarteten Breitensuche der Tiefe k induziert wird. N. Alon hat gezeigt, dass für jedes $\epsilon > 0$ und jeden durch d im Grad beschränkten Graph G ein Graph H mit einer von G unabhängigen Größe existiert, dessen Verteilung von k-Disks sich in der ℓ_1-Norm höchstens um ϵ von der Verteilung der k-Disks in G unterscheidet. In dieser Arbeit wird eine explizite obere Schranke in $\mathcal{O}(d^{3k+2}2^{7.5d^k}/\epsilon^4)$ für $|H|$ für den Fall gezeigt, dass G planar ist, und dieses Ergebnis auf hyperfinite Graphen verallgemeinert. Darüber hinaus werden weiterführende Resultate über k-Disks und k-Disk-Verteilungen vorgestellt.

Abstract

Given an undirected graph G, the k-disk of a node u is the rooted subgraph that is induced by a breadth-first search of maximum depth k that is started at u. It has been proved by N. Alon that, for every $\epsilon > 0$ and every degree-bounded graph G, there exists a graph H whose size is independent of the size of G such that the ℓ_1 distance between the distributions of k-disks in G and H differs only by ϵ. In this thesis, we give an explicit upper bound of order $\mathcal{O}(d^{3k+2}2^{7.5d^k}/\epsilon^4)$ on the size of H if G is a planar graph with degree bounded by d. Furthermore, we generalize this result to hyperfinite graphs and provide some secondary results on k-disks and k-disk distributions.

Contents

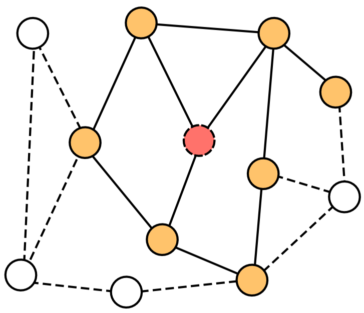
1. Introduction

Obtaining useful information about large datasets is one of the challenging tasks in recent applications of computer science. Many of the datasets that appear in practice can be thought of as graphs. Typical examples of such data are relations in social networks, recordings of browsing behavior on websites and the webgraph. These datasets often are too large to be processed by polynomial- or even linear-time algorithms, thus sublinear algorithms are desirable in these cases. The framework of property testing provides a way to relax traditional decision problems. Often, these relaxed problems can be solved in sublinear time.

In property testing, one has to decide whether a given input (e.g., an image of a leaf) has a property (e.g., is red) or is *far away* from having it (e.g., not red nor orange) by only looking at a small, randomly chosen portion of it (e.g., 100 pixels) such that the answer is correct with constant probability. More technically, an input is said to be ϵ-far from a property if one has to change more than an ϵ-fraction of its representation to obtain an object that has the property. Otherwise the input is ϵ-close to the property. Therefore, the representation model of an object is closely related to what can be tested with a limited amount of effort.

Many results that have been obtained so far hold for the dense graph model by Goldreich et al. [27]. In this model, a property testing algorithm can query whether two nodes u and v are adjacent. Property testing in (undirected) dense graphs has been studied quite successfully. For example, Alon et al. [7] have given a combinatorial characterization of the properties that can be tested by a sequence of queries whose length depends on the proximity parameter ϵ only.

Another major model in graph property testing is the bounded degree graph model by Goldreich and Ron [24], where every node has at most d neighbors. The number of edges in a degree-bounded graph on n nodes is at most $dn/2 = \mathcal{O}(n)$, which is significantly smaller than the maximum $\mathcal{O}(n^2)$ for arbitrary graphs (without multiple edges), and therefore such graphs are called sparse. Consequently, a degree-bounded graph is ϵ-far from a property if at least ϵdn edges have to be changed to obtain a graph that has the property. A property testing algorithm for degree-bounded graphs is given the size of the input graph and its maximum node degree d. It can access the graph by querying an oracle for the j-th neighbor of the i-th node. This way the testing algorithm is restricted to choosing a small number of nodes and their neighborhoods for exploration. Typical graph property tests ask for bipartiteness, connectivity, k-colorability or triangle-freeness of a graph. A popular tool to explore the input graph are random walks starting at the sampled nodes (e.g., for testing bipartiteness [25] or expansion [16, 26]). Another frequently used technique is to examine the neighborhood of every sampled node down to a certain depth (e.g., for testing connectivity or cycle-freeness [24]). This can be done by performing a breadth-first search of depth k at a sampled node and considering the subgraph induced by the discovered nodes. This subgraph, which is rooted at the start node r and is induced by all nodes with a distance of at most k to r, is called the *k-disk* of r.



2-disk of the red node (dashed circle). Orange nodes and solid edges belong to the k-disk.

The k-disks of all nodes of a graph G can be used to construct a feature vector for G that characterizes its local structure. Each entry in the vector corresponds to a maximal set of isomorphic k-disks, and its value is the number of occurrences in the graph, i.e., the number of nodes with a k-disk of this isomorphism type. We call this vector the *k-disk vector* of the graph. Since the node degree is bounded by a constant d, the number of isomorphism classes of k-disks is finite, and therefore the dimension of the vector is finite, too. If all entries of a k-disk vector are divided by the size of the graph, we call it a *normalized k-disk vector*.

Lovász raised the following question regarding a degree-bounded graph G of arbitrary size and its normalized k-disk vector [cf. 33, Question 7]:

> Given a degree-bounded graph G of size n, is there a graph H of constant size whose normalized k-disk vector is similar to that of G?

The similarity of two vectors is measured by their ℓ_1 distance here, and the size of H is constant in terms of being independent of the size of G (see Theorem 3.1 on page 28 for the formal statement). As a matter of fact, Alon has proved the existence of such a small graph H for every degree-bounded graph G (cf. proof recited by Lovász [42, Proposition 19.10] and Theorem 3.1). Unfortunately, there does not follow any effective bound on the size of H.

In this thesis, we prove the following explicit upper bound on the size of H if G is planar (see Corollary 3.4 on page 30 for the precise statement):

Theorem. *Let G be a planar graph with degree bounded by d, let $k \geq 0$ be an integer and $\epsilon > 0$. There exists a planar graph H of size at most $10^6 \cdot d^{3k+2} 2^{7.5d^k} / \epsilon^4$ such that the ℓ_1 distance between the normalized k-disk vectors of G and H is at most ϵ.*

The family of planar graphs is a subclass of the class of *hyperfinite* graphs, i.e., graphs that can be split into connected components of constant size by removing a constant fraction of edges only. Actually, we obtain a bound for arbitrary families of hyperfinite graphs and use it to derive explicit bounds on the size of H if G is a forest or a planar graph or does not contain a fixed minor, respectively (see Theorem 3.2 on page 29 and Corollaries 3.3 to 3.5 for the precise statements).

In the remainder of this chapter, we present some modest examples as a gentle introduction to the formal problem. Besides, we give an overview of previous work related to the definitions and tools that come into play when proving explicit bounds for hyperfinite graphs later. We end this chapter with some abbreviations and notation.

In Chapter 2, the definitions that are necessary to formally state the results and their proofs are introduced. The main results are formally introduced and proved in Chapter 3. Since k-disks describe the local structure of a graph, one can encode a graph as its k-disks vector plus some edge operations as global *fill-in* graph. We give a lower bound on k when a planar graph is encoded this way. In Chapter 4, we prove some results related to the feasibility of k-disk vectors and the relation between expander graphs and their k-disk vectors. Furthermore, we present an empirical study of the k-disks of real world networks. Note that the index at the end of this thesis provides quick access to the definitions of terms and symbols.

1.1. Example

Before we give two simple examples of families of arbitrarily large graphs where a small graph with similar normalized k-disk vector can be constructed easily, let us consider a single graph G first. If we know the k-disk vector of G, we know the structure of each node's neighborhood in G. It is tempting to conjecture that we can *stick* these neighborhoods *together* to recover G. However, without more information there is no way to ensure that we retrieve G. There may exist many non-isomorphic graphs of equal size that share the same k-disk vector. Consider, for example, the graph in Fig. 1.1 on page 9: If we swap the orange and the red (dashed border) subgraph, the k-disk vector of the graph does not change because all (sub-)paths of length k look the same and prevent two knots from being in the same k-disk. However, this graph is not isomorphic to the original graph.

On the positive side, there exists a connection between the k-disk vectors of two hyperfinite graphs and their global structure as shown by Newman and Sohler [44]. Roughly speaking, their result states that for two sufficiently large, $\varphi(\mu)$-hyperfinite graphs (see Definition 2.12 on page 19) G_1, G_2 and $\epsilon > 0$ there exist $k := k(\epsilon, d, \varphi(\cdot))$ and $\delta := \delta(\epsilon, d, \varphi(\cdot))$ such that G_1 and G_2 are ϵ-close if the ℓ_1 distance of their normalized k-disk vectors differs by at most δ.

In Chapter 3 we present a construction that yields, for every degree-bounded, planar graph G of arbitrary size and fixed $\epsilon > 0$, a planar graph H of constant size whose ℓ_1 distance to the input graph's normalized k-disk vector is only ϵ. This allows us to estimate the size of H. In the following, we illuminate the basic idea by examples. We consider two basic families of planar graphs and construct a small graph H with a similar normalized k-disk vector for each of them in the following. Although both families contain a graph G of size at least n for every $n \in \mathbb{N}$, the small graph H is always the same (within the family), and its size depends on the proximity parameter ϵ only, i.e., the desired maximum ℓ_1 distance of the normalized k-disk vectors of H and any G from the family.

First Example: Cycle Graphs

Let G be an n-node graph that consists of one cycle and let $k = 2$. For the sake of simplicity, assume that $n > 5$. In this case, all nodes of G share the same type of k-disk, which is not the whole graph (see Fig. 1.2). We cut a path P_ℓ of *small* length ℓ, say $4 < \ell \ll n$ for now, out of G as a candidate for the small graph H (i.e., H is a chain of $\ell + 1$ nodes). There are only six isomorphism types of 2-disks that can be centered at nodes of P_ℓ: they are shown in Fig. 1.3. We assume that a 2-disk vector indicates the number of 2-disks isomorphic to type (a), (b), (c), (d), (e) and (f) in this order (see Fig. 1.3). All other types of k-disks are omitted as they can never show up in this example anyway. Then, the 2-disk vector of P_ℓ equals $\mathrm{dist}_2(P_\ell) = (0, 0, 2, 0, 2, \ell - 3)^T$, while the 2-disk vector of the original graph G equals $\mathrm{dist}_2(G) = (0, 0, 0, 0, 0, n)^T$. The ℓ_1 distance of the normalized 2-disk vectors of G and P_ℓ depends on ℓ:

$$
\begin{aligned}
& ||\mathrm{freq}_2(G) - \mathrm{freq}_2(P_\ell)\,||_1 \\
= {} & \left|\left|\Big(0, 0, 0, 0, 0, 1\Big)^T - \Big(0, 0, \frac{2}{\ell+1}, 0, \frac{2}{\ell+1}, 0, \frac{\ell-3}{\ell+1}\Big)^T\right|\right|_1 \\
= {} & \frac{2 + 2 + 4}{\ell + 1} = \frac{8}{\ell + 1}\,.
\end{aligned}
$$

In other words, we can guarantee that $||\mathrm{freq}_2(G) - \mathrm{freq}_2(H)\,||_1 \leq \epsilon$ by choosing $H := P_\ell$, where $\ell := \lceil 8/\epsilon - 1 \rceil \leq 8/\epsilon = \mathcal{O}(1/\epsilon)$.

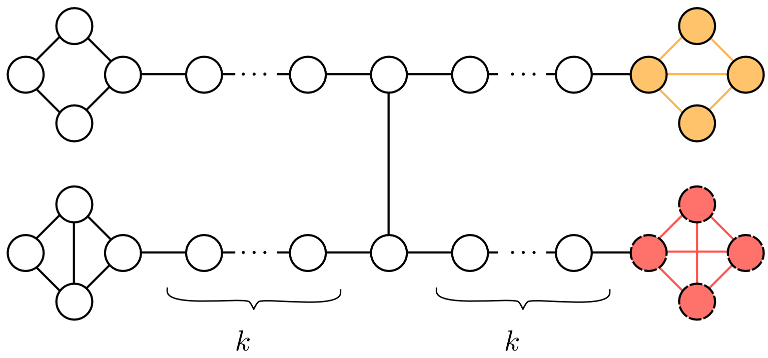


Figure 1.1.: The graph above and the graph where the orange and the red (dashed border) subgraph are swapped have the same k-disk vector but are not isomorphic.

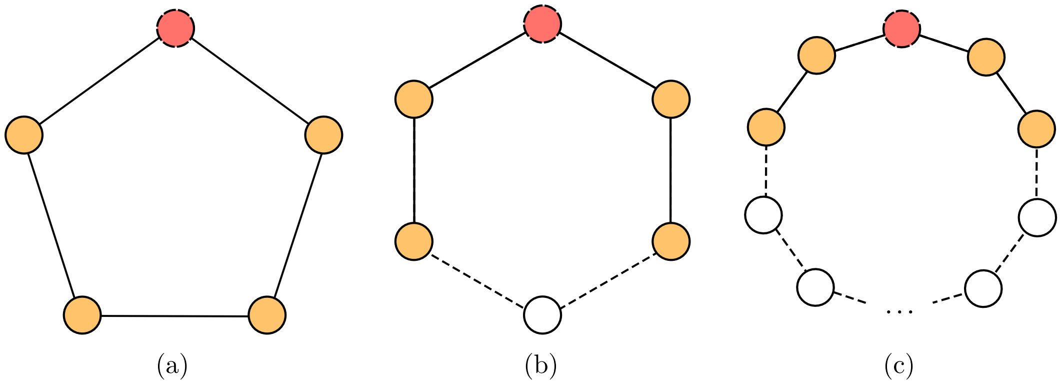


Figure 1.2.: 5-, 6- and n-node cycle graphs. A 2-disk rooted at the red node (dashed circle) is shown for each graph. Colored nodes and solid edges belong to the k-disk, white nodes and dashed edges belong to the graph only. The normalized 2-disk vectors of (b) and (c) are equal. All 2-disks of (a) are cycles itself and differ from the 2-disks of (b) and (c), which are simple paths of length 4.

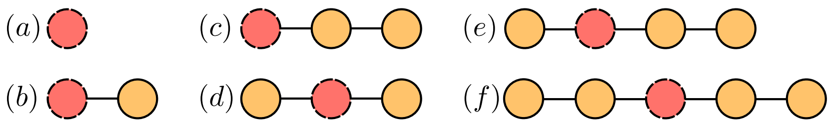


Figure 1.3.: All six 2-disk isomorphism types that can be centered at the nodes of a path. The roots are colored red (dashed circles).

Remark We note that in this example, one can actually obtain an optimal graph by constructing a cycle of length six. However, as this graph is not a subgraph of the original graph, it does not resemble the idea of the main result's proof in Chapter 3.

If all nodes of a graph have degree at most two, the graph contains paths and cycles only. One can extend this example to these graphs and arbitrary k by combining paths and cycles of length at most k and (possibly multiple copies of) a suitably longer path. We proceed to the next example instead and deal with this scenario later as a special case of graphs where all nodes have degree at most three.

Although cycle graphs are not trees, they are very *treelike*. In fact, every cycle graph of arbitrary size has treewidth* two. The treewidth of planar graphs is, however, unbounded in general. Hence, the family of planar graphs we consider in the next example has unbounded treewidth.

Second Example: Grid Graphs

An $n \times n$ grid graph is a graph $G = (V, E)$ with node set $V = \{(i, j) \mid i, j \in \{1, \dots, n\}$ where each node $(i, j) \in V$ is adjacent to $\{(i-1, j), (i+1, j), (i, j-1), (i, j+1)\} \cap V$. It has treewidth n [cf. 9], and therefore the maximum treewidth of a graph from the family of all $n \times n$ grids is unbounded.

Let G be an $n \times n$ grid graph where $n \geq 2$, and let $k = 1$. There are three possible 1-disks that can be centered at a node of an $n \times n$ grid graph (see Fig. 1.4): a root node with two, three or four adjacent nodes, respectively. In particular, all four nodes in the corners have two neighbors, all $4n - 8$ other nodes on the border have three neighbors, and all $(n-2)^2$ inner nodes have four neighbors. We assume that a 1-disk vector indicates the number of 1-disks in the order of their size. Then, the 1-disk vector of an $n \times n$ grid graph equals $\left(4, 4n-8, (n-2)^2\right)^T$. We cut an $\ell \times \ell$ grid graph G_ℓ, say $2 \leq \ell \ll n$ for now, out of G and compare its normalized 1-disk

*Definition [28, 46]: A tree composition of a graph $G = (V, E)$ is a tree $T = (X, Y)$ whose nodes $X_1, \dots, X_m$ are subsets of V and that satisfies the following conditions: *a)* $\bigcup_i X_i = V$ *b)* $\forall (u, v) \in E : \exists i : \{u, v\} \subseteq X_i$ *c)* $\forall i, j, k : X_k$ is on the path from X_i to $X_j \Rightarrow X_i \cap X_j \subseteq X_k$. The width of a tree decomposition is the size of its largest node(s) minus one, i.e., $\max_i |X_i| - 1$. The treewidth of a graph G is the minimum width among all tree decompositions of G. Trees have treewidth one (or zero if they contain no edge). All other graphs have a larger treewidth.

vector to that of G:

$$
\begin{aligned}
&||\mathrm{freq}_1(G) - \mathrm{freq}_1(G_\ell)\,||_1 \\
&= \left|\left|\left(\frac{4}{n^2}, \frac{4n-8}{n^2}, \frac{(n-2)^2}{n^2}\right)^T - \left(\frac{4}{\ell^2}, \frac{4\ell-8}{\ell^2}, \frac{(\ell-2)^2}{\ell^2}\right)^T\right|\right|_1 \\
&\leq \left|\frac{4}{n^2} - \frac{4}{\ell^2}\right| + \left[\left|\frac{4}{n} - \frac{4}{\ell}\right| + \left|\frac{8}{n^2} - \frac{8}{\ell^2}\right|\right] + \left[\left|\frac{n^2}{n^2} - \frac{\ell^2}{\ell^2}\right| + \left|\frac{4}{\ell} - \frac{4}{n}\right| + \left|\frac{4}{n^2} - \frac{4}{\ell^2}\right|\right] \\
&\leq \left|\frac{4}{\ell^2}\right| + \left|\frac{4}{\ell}\right| + \left|\frac{8}{\ell^2}\right| + |0| + \left|\frac{4}{\ell}\right| + \left|\frac{4}{\ell^2}\right| \leq \frac{16}{\ell^2} + \frac{8}{\ell} \leq \frac{16}{\ell}
\end{aligned}
$$

We can guarantee that $||\mathrm{freq}_k(G) - \mathrm{freq}_k(H)\,||_1 \leq \epsilon$ by choosing $H := G_\ell$, where $\ell := \left\lceil \frac{16}{\epsilon} \right\rceil \leq \frac{16}{\epsilon} + 1 = \mathcal{O}(1/\epsilon)$.

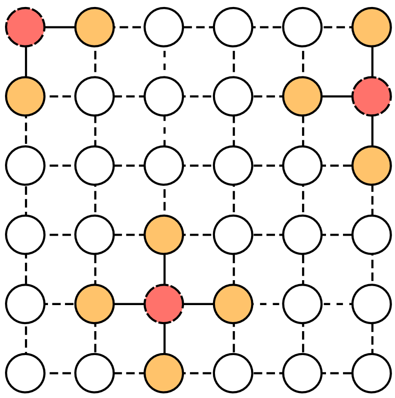


Figure 1.4.: A 6×6 grid graph and its three 1-disk isomorphism types. Colored nodes and solid edges belong to the 1-disks. White nodes and dashed edges belong to the graph only. The k-disks' roots are colored red (dashed circles).

The preceding two examples illustrate the basic idea of constructing a small graph H with similar k-disk vector from a hyperfinite graph G: A small component H of the original graph G captured enough of its characteristics to function as an approximation to the k-disk vector of G. In particular, the graph H was some kind of a *small-scale version* of G. Therefore, the normalized k-disk vectors of G and H were very similar. However, not every hyperfinite graph contains a single connected component like H. In Chapter 3 we will generalize the concept by combining (possibly) more than one component to construct H for any degree-bounded, hyperfinite graph.

1.2. Related Work

For dense graphs, the regularity lemma by Szemerédi [48] is a powerful tool to get a small characterization of an arbitrary graph. It states that the nodes of any graph G can be partitioned into a constant number of sets of (nearly) equal size such that the edges between all but a few pairs of these sets behave quasirandomly. The inner structure of a partition set is negligible for many purposes if G is dense, and therefore examining (the densities of) the edges between the partitions is enough to, e.g., test certain properties of G. There exist several alternate formulations as well as strong and weak versions of the regularity lemma, which can be chosen to trade the lemma's practicality for a stronger conclusion and vice versa. The proposition on page 6 that was proved by Alon can be seen as an analogue for sparse graphs to a weak version of the regularity lemma for dense graphs, which was derived by Lovász [41, Lemma 5.3] from a result by Frieze and Kannan [22]. It was also put up as a finite version of the Aldous-Lyons conjecture from the theory of graph limits [1] by Lovász.

In this thesis, we focus on hyperfinite graphs, which were introduced by Elek [19]. It was shown by Newman and Sohler [44] that two degree-bounded hyperfinite graphs of equal size can be transfered into each other by only changing a constant fraction of their edges if their normalized k-disk vectors are similar. Their result also yields a property tester that decides whether two hyperfinite graphs are isomorphic or far from being isomorphic. It is a generalization of a result by Benjamini et al. [8], who proved that every minor-closed property (i.e., if G has the property, then so does every minor of G) is testable. Another way of dealing with sparse graphs, which was proposed by Elek and Lippner [20], is to partition them into *highly homogeneous parts*. Roughly speaking, for every part P in this partitioning, the normalized k-disk vector of a subgraph S is similar to the part's k-disk vector if S is induced by sufficiently many nodes and the cut between S and $P \setminus S$ is small.

As mentioned before, there exist local views of sparse graphs other than the k-disk approach. Huang et al. [32] studied the 3-local profiles of arbitrary graphs, i.e., the probability vector for the events that three randomly chosen nodes span $(0, 1, 2, 3)$ edges. Given an arbitrary graph, one can identify the isomorphism type of every subgraph induced by exactly m nodes and count their frequencies in a vector similar to a k-disk vector. Ugander et al. [50] gave constraints for the set of all feasible frequency vectors and investigated their occurrence in social network graphs. One might expect that some of these subgraphs are characteristic of graphs from certain applications and appear more often than others. These subgraphs are called motifs and were studied by Milo et al. [43] in various types of real world instances.

1.3. Abbreviations and Notation

From now on, we silently assume that all variables named d, k, n or N are restricted to integer values. All other variables may attain real values if not specified otherwise.

Bachmann-Landau notation – or simply asymptotic notation – is often used to describe the limiting behavior of a function. The notation we will use for asymptotic upper bounds on a function $g(x)$ as x tends to infinity is given by the following definition.

Definition 1.1 (Big-O notation). Let $f, g : \mathbb{R}^N \to \mathbb{R}$ be two functions. Then, $g(\vec{\boldsymbol{x}}) \in \mathcal{O}(f(\vec{\boldsymbol{x}}))$ iff there exist $c > 0$ and $x \in \mathbb{R}$ such that for every $x_1, \ldots, x_d \geq x$ it holds that $g(x_1, \ldots, x_d) \leq cf(x_1, \ldots, x_d)$. ■

We note that, if more than one variable is involved, i.e., $N > 1$, some properties known from asymptotic calculus for functions on one variable do not hold. Especially, one should not add, multiply or otherwise *join* asymptotic expressions with multiple variables in further calculations. Therefore, we avoid asymptotic calculus in this thesis and derive asymptotic bounds only as an end result. See [31] for more details and other possible definitions of asymptotic notation with multiple variables.

Definition 1.2 (Little-o notation). Let $f, g : \mathbb{R} \to \mathbb{R}$ be two functions. Then, $g(x) \in o(f(x))$ iff $\lim_{x\to\infty} |g(x)/f(x)| = 0$. ■

For any binary relation $\sim$ on X and two sets $A, B \subseteq X$, $A \sim B$ is defined as $a \sim b$ for every $a \in A$ and $b \in B$. Given two sets X, Y and a function $f : X \to Y$, we define $f(X') := \{f(x) \mid x \in X'\}$ for every subset $X' \subseteq X$ and $f((x_1, \ldots, x_\ell)) = (f(x_1), \ldots, f(x_\ell))$ for every $x \in X^\ell$. Instead of writing $\{1, 2, \ldots, n\}$, we use $[n]$ as an abbreviation. The ℓ_p norm of a vector $\vec{\boldsymbol{x}}$ is denoted by $||\vec{\boldsymbol{x}}||_p := \sqrt[p]{\sum_i |x_i|^p}$ for $p \geq 1$. It induces the ℓ_p distance $||\vec{\boldsymbol{x}} - \vec{\boldsymbol{y}}||_p := \sqrt[p]{\sum_i |x_i - y_i|^p}$ of two vectors $\vec{\boldsymbol{x}}$ and $\vec{\boldsymbol{y}}$.

2. Preliminaries

This chapter covers all definitions and notation that are needed to state the problem and the results later. Some of them have already been used in the previous chapter and are now introduced formally.

2.1. Graph Theory and Carathéodory's Theorem

In this section, we recall basic definitions from graph theory and introduce k-disks, the family of hyperfinite graphs, which are required to state the formal problem and to prove the main results, and Carathéodory's Theorem, the proof's main technical tool.

2.1.1. Basic Definitions

Definition 2.1 (Graph). A tuple $G := (V, E)$ where V is a set of nodes (or vertices) and $E \subseteq \{\{u, v\} \mid u, v \in V \wedge u \neq v\}$ is a set of edges is called an **(undirected) graph** without loops and multiple edges. We denote the vertices of an arbitrary graph G' by $V(G')$ and the edges by $E(G')$. ■

Let $G = (V, E)$ be a graph. The **size** of G is $|G| := |V(G)|$, i.e., the number of its nodes. Two nodes $u, v \in V$ are **adjacent** iff $\{u, v\} \in E$. The **adjacency matrix** of a graph $G = (V, E)$, $V = \{v_1, \ldots, v_n\}$, is an $n \times n$ matrix where the entry (i, j) equals one iff $\{v_i, v_j\} \in E$ and it equals zero otherwise. The **neighboorhood** of a node u is the set of adjacent nodes $\Gamma(u) := \{v \mid \{u, v\} \in E\}$. An edge $\{u, v\} \in E$ and a node $w \in V$ are **incident** iff $w \in \{u, v\}$. The **degree** of a node is the number of edges that are incident to it. We say that a graph is ***d*-bounded** iff its maximum node degree is bounded by d. It is ***d*-regular** iff all nodes have degree of exactly d.

The **boundary** ∂S of a subset of nodes $S \subseteq V$ is the set of edges with exactly one node in S, i.e., $\partial S := E \cap \{\{u, v\} \mid u \in S \wedge v \in V \setminus S\}$. A **cut** is a set of edges that partitions the nodes of a graph into two sets. A tuple (G, r) where G is a graph and $r \in V(G)$ is called **rooted graph** with root r.

Two graphs $G = (V_1, E_1)$ and $H = (V_2, E_2)$ are **isomorphic**, written as $G \cong H$, iff $|V_1| = |V_2|$ and there exists a bijection $f : V_1 \to V_2$ such that $\{v, w\} \in E_1 \Leftrightarrow \{f(v), f(w)\} \in E_2$.

It is common to denote the number of nodes by n and the number of edges by m. Undirected edges are often written as (u, v) instead of $\{u, v\}$. To simplify notation, we occasionally write $u \in G$ instead of $u \in V(G)$ and $(u, v) \in G$ instead of $(u, v) \in E(G)$. Given two graphs $G = (V, E)$ and $G' = (V', E')$, we write $G' \subseteq G$ to state that G' is a subgraph of G, i.e., $V' \subseteq V$ and $E' \subseteq E$. In the following, we assume that the set of nodes is finite unless stated otherwise.

Taking only a subset V' of nodes and the edges between them into account, one obtains the subgraph induced by V'.

Definition 2.2 (Induced subgraph). Given a graph $G = (V, E)$ and a subset $V' \subseteq V$, the subgraph $G[V']$ induced by V' is defined as

$$\begin{aligned} G[V'] &:= (V', E') \\ E' &:= \{(u, v) \mid u, v \in V' \wedge (u, v) \in E\} . \end{aligned} \qquad \blacksquare$$

Multiple graphs can be joined in a copy-and-paste manner.

Definition 2.3 (Joined graphs). Let $G_1 = (V_1, E_1), \ldots, G_\ell = (V_\ell, E_\ell)$ be graphs where $V_i = \{v_{i,1}, \ldots, v_{i,|V_i|}\}$. The graph $\mathrm{join}(G_1, \ldots, G_\ell)$ obtained by joining $G_1, \ldots, G_\ell$ is defined as

$$\begin{aligned} \mathrm{join}(G_1, \ldots, G_\ell) &:= \left(\bigcup_{i=1}^{\ell} f(V_i), \bigcup_{i=1}^{\ell} f(E_i) \right) \\ f(v_{i,j}) &:= u_p \text{ where } p := \sum_{q=1}^{i-1} |V_q| + j . \end{aligned} \qquad \blacksquare$$

In particular, it is possible to join multiple copies of the same graph and obtain a graph that contains the original graph multiple times (as distinct induced subgraphs).

A **path** of length ℓ is an alternating sequence of nodes and edges $u_1, (u_1, u_2), u_2, \ldots, (u_{\ell-1}, u_\ell), u_\ell$ where $u_i \in V$ and $(u_i, u_{i+1}) \in E$. It is sufficient to state the nodes or the edges only to define a path. The length of the **shortest path** between two nodes u and v is denoted by $\mathrm{d}_G(u, v)$ – or $\mathrm{d}(u, v)$ for short. A graph is **connected** iff there exists a path between each pair of nodes. A **connected component** of a graph is a connected induced subgraph that cannot be enlarged by adding nodes to the inducing subset V' while staying connected.

A path of length at least three that starts and ends at the same node and visits each node in between only once is called **cycle**. The **girth** of a graph is the length of a shortest cycle and is denoted by $\mathrm{girth}(G)$. Paths without cycles are **simple**.

The **complete graph**, or **clique**, on t nodes, i.e., (V, E) where $|V| = t$ and $E = V \times V \setminus \{(u, u) \mid u \in V\}$, is denoted by K_t. Similarly, the **complete bipartite graph** on $s + t$ nodes, i.e., $(V_1 \cup V_2, E)$ where $|V_1| = s$, $|V_2| = t$ and $E = V_1 \times V_2 \setminus \{(u, u) \mid u \in V_1 \cup V_2\}$, is denoted by $K_{s,t}$.

2.1.2. k-Disks and k-Disk Vectors

Starting from an arbitrary node u_0 of a graph G, one can examine its local surroundings by looking at the subgraph induced by all nodes *near* to u_0. If near nodes are defined as those that can be reached from u_0 by a path of length at most k, the subgraph induced by u and these near nodes is called k-disk. To put it another way, the k-disk of a node u_0 is the subgraph induced by all nodes that are discovered by a breadth-first search of depth k starting at u_0. See Fig. 2.1 for an example of a 2-disk.

Definition 2.4 (k-disk). Given a graph $G = (V, E)$, a node $u_0 \in V$ and a non-negative integer k, the k-disk centered at u_0 is a rooted subgraph $G[V']$ with root u_0 and is denoted by $\mathrm{disk}_k(G, u_0)$. The set V' consists of all nodes u_ℓ with distance between u_0 and u_ℓ at most k.

$$\begin{aligned}\mathrm{disk}_k(G, u_0) &:= (G[V'], u_0).\\ V' &:= \{u \mid u \in V \,\wedge\, \mathrm{d}_G(u_0, u) \leq k\}\,.\end{aligned}$$

Two k-disks $D_1 := ((V_1, E_1), u_1)$ and $D_2 := ((V_2, E_2), u_2)$ are isomorphic iff there exists a bijection $f : V_1 \to V_2$ such that (V_1, E_1) and (V_2, E_2) are isomorphic and f maps the root of D_1 to the root of D_2, i.e.,

$$\begin{aligned}(v, w) \in E_1 &\Leftrightarrow (f(v), f(w)) \in E_2\\ f(u_1) &= u_2\,.\end{aligned}$$

We write $D_1 \cong D_2$ to state that D_1 and D_2 are isomorphic. For each k-disk D, its isomorphism type is defined as the equivalence class that is induced by the isomorphism relation, i.e., $[D]_{\cong} = \{D' \mid D \cong D'\}$. Given a family $\mathcal{F}$ of graphs, we denote the set of representatives of all equivalence classes of k-disks that appear in at least one d-bounded graph from $\mathcal{F}$ by $R_{\mathcal{F}}(d, k)$. ■

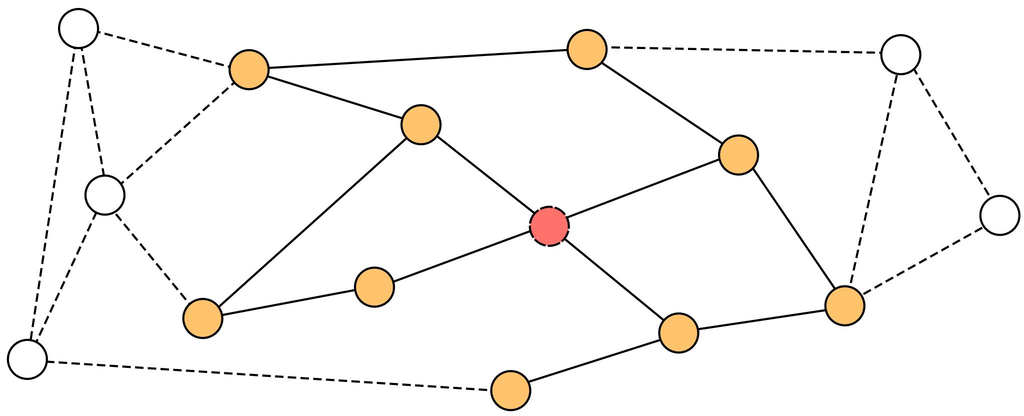


Figure 2.1.: A 2-disk rooted at the red node (dashed circle) of the shown graph. Colored nodes and solid edges belong to the 2-disk, white nodes and dashed edges belong only to the rest of the graph.

For every non-negative integer k and every integer d, the number of isomorphism types of k-disks with maximum node degree d is finite. Given a d-bounded graph G, one can determine the type of k-disk for every node of G and count the number of occurrences of each type using a check sheet. This check sheet can be formalized as the k-disk vector of G.

Definition 2.5 (k-disk vector). Let $\mathcal{F}$ be a family of graphs and let $N := N_{\mathcal{F}}(d,k) := |R_{\mathcal{F}}(d,k)|$, i.e., the number of isomorphism types $\mathcal{T}_1, \ldots, \mathcal{T}_{|R_{\mathcal{F}}(d,k)|}$ of the set of k-disks that appear in at least one d-bounded graph from $\mathcal{F}$. Given a graph $G \in \mathcal{F}$ and a non-negative integer k, a k-disk vector of G is a vector indexed by $\mathcal{T}_1, \ldots, \mathcal{T}_N$ in an arbitrary but fixed order. The (unnormalized) k-disk vector $\mathrm{dist}_k(G)$ counts, for each isomorphism type $\mathcal{T}_i$, the number of k-disks of G that are isomorphic to $\mathcal{T}_i$.

$$\mathrm{dist}_k(G) := \begin{pmatrix} \sum_{u \in V} \mathcal{I}_G(u, \mathcal{T}_1) \\ \vdots \\ \sum_{u \in V} \mathcal{I}_G(u, \mathcal{T}_N) \end{pmatrix}$$

$$\mathcal{I}_G(u, \mathcal{T}_i) := \begin{cases} 1 & \text{if } \mathrm{disk}_k(G,u) \cong \mathcal{T}_i \\ 0 & \text{otherwise}\,. \end{cases}$$

The normalized k-disk vector $\mathrm{freq}_k(G)$ counts, for each isomorphism type $\mathcal{T}$, the fraction of k-disks of G that are isomorphic to $\mathcal{T}$.

$$\mathrm{freq}_k(G) := \frac{\mathrm{dist}_k(G)}{||\mathrm{dist}_k(G)||_1} .$$

We write $\mathrm{freq}_k(G, \mathcal{T})$ to denote the entry of $\mathrm{freq}_k(G)$ that corresponds to the fraction of k-disks that are isomorphic to $\mathcal{T}$. ■

Since every node in a graph maps to exactly one k-disk, the number of nodes in a graph G equals the ℓ_1 norm of its k-disk vector, i.e., $|V(G)| = ||\mathrm{dist}_k(G)||_1$. The dimension of the space spanned by all normalized k-disk vectors is at most $N-1$ because for every normalized k-disk vector $(x_1, \ldots, x_N)^T$ it holds that $x_1 + \ldots + x_{N-1} - 1 = -x_N$.

Definition 2.6. Let $\mathcal{F}$ be a family of graphs and let d, k be non-negative integers. Denote the set $\mathcal{F}$ restricted to d-bounded graphs by $\mathcal{F}_d$. Then, $M(\epsilon) := M_{\mathcal{F}}(\epsilon, d, k)$ denotes the minimum size such that, for every graph $G \in \mathcal{F}_d$, there exists a graph $H \in \mathcal{F}_d$ of size at most $M_{\mathcal{F}}(\epsilon, d, k)$ such that its normalized k-disk vector's ℓ_1 distance to the normalized k-disk vector of G is at most ϵ.

$$M(\epsilon) := M_{\mathcal{F}}(\epsilon, d, k) := \max_{G \in \mathcal{F}_d} \min \left\{ |H| \;\middle|\; H \in \mathcal{F}_d \wedge ||\mathrm{freq}_k(G) - \mathrm{freq}_k(H)||_1 \leq \epsilon \right\} . \quad ■$$

We recall from Chapter 1 that Alon has proved that $M_{\mathcal{F}}(\epsilon, d, k)$ is well-defined and finite for the family of all graphs and every choice of $\epsilon > 0$ and $d, k \geq 0$ (cf. Theorem 3.1 on page 28).

2.1.3. Forests, Planar and Minor-Free Graphs

Forests are graphs that contain no cycles. Trees are connected forests.

Definition 2.7 (Trees and Forests). Let T be a connected graph that contains no cycle. Then, T is a tree. A tree $T = (V, E)$ can be represented as a recursive structure by rooting it at one of its nodes $u \in V(T)$.

$$T(u) := T_V(u) := (u, \{T_{V \setminus \{u\}}(v) \mid (u, v) \in E(T) \wedge v \in V\}) .$$

We call u the root of $T(u)$. Let $T_1, \ldots, T_\ell$ be trees. Then, $\mathrm{join}(T_1, \ldots, T_\ell)$ is a **forest**. ■

The family of graphs that can be embedded in the plane without crossing edges is called planar graphs.

Definition 2.8 (Planar Graph). A graph $G = (V, E)$ is planar iff there exists an injective function $f : V \to \mathbb{R}^2$ and for every edge $e = (u, v) \in E$ there exists a continuous function $g_e : [0, 1] \to \mathbb{R}^2$ such that it holds that

$$\begin{aligned} &g_e(0) = f(u) \text{ and } g_e(1) = f(v) && \forall\, e = (u, v) \in E \\ &g_e(x) \neq g_f(y) && \forall\, e, f \in E, e \neq f,\ x, y \in (0, 1)\,. \end{aligned}$$ ■

Given a graph G, melting to adjacent nodes u and v together is called an edge contraction.

Definition 2.9 (Edge contraction). Let $G = (V, E)$ be a graph and $(u, v) \in E$ be an edge. Let $f : V \to V \setminus \{u, v\} \cup \{w\}$ be a function that maps u and v to w and all other nodes to themselves. Contracting the edge (u, v) results in the graph $G' = (V', E')$ where $V' := \{f(x) \mid x \in V\}$ and $E' := \{(f(x), f(y)) \mid (x, y) \in E \setminus (u, v)\}$. ■

Informally, a graph G contains another graph M as a minor iff the nodes of M can be mapped to nodes of G such that all existing paths in M are preserved.

Definition 2.10 (Minor). Let G and M be graphs. The graph G contains M as a minor iff one can obtain a graph $G' \cong M$ from G by contracting edges and deleting nodes and edges of G. Conversely, G is M-free iff it does not contain M as a minor. ■

Forests are cycle-free, i.e., K_3-free. Kuratowski [36] has proved that planar graphs are K_5-free and $K_{3,3}$-free.

Theorem 2.11 (Kuratowski's Theorem [36]). *A graph G is planar iff it contains neither K_5 nor $K_{3,3}$ as a minor.*

2.1.4. Hyperfinite Graphs

Hyperfinite graphs were first introduced by Elek [19]. Informally, a graph is called hyperfinite iff one can remove a small constant fraction of all of its edges in such a way that the graph falls apart into connected components of constant size.

Definition 2.12 (Hyperfinite graph [19, 29]). Let $G = (V, E)$ be a graph and let $\mu > 0$. Then G is called (μ, φ)-hyperfinite iff one can remove $\mu \cdot |V|$ edges from G and obtain a graph whose connected components have size at most φ. For a function $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$, a family of graphs is called $\varphi(\mu)$-hyperfinite iff, for every $\mu > 0$, every graph in the family is $(\mu, \varphi(\mu))$-hyperfinite. ■

Some major families of degree-bounded graphs are known to be hyperfinite. A basic family of graphs that satisfies Definition 2.12 is the family of forests.

Theorem 2.13 (Forests are hyperfinite [cf. 40]). *Let $G = (V, E)$ be a forest of size n with maximum node degree d. Then, G is $\varphi(\mu)$-hyperfinite, where $\varphi(\mu) = 3d/\mu$.*

Another family that is also known to be hyperfinite is the family of planar graphs.

Theorem 2.14 (Planar graphs are hyperfinite [cf. 40]). *Let $G = (V, E)$ be a planar graph of size n with maximum node degree d. Then, G is $\varphi(\mu)$-hyperfinite, where $\varphi(\mu) = \gamma^2[(3+\sqrt{6})d/\mu]^2$ and $\gamma \leq 2\sqrt{2}$.*

Given a graph M, one can define the set of all graphs that do not contain M as a minor. It is known that, for every fixed minor M, this family is hyperfinite, too.

Theorem 2.15 (Minor-free graphs are hyperfinite [cf. 6, 35]). *Let $G = (V, E)$ be a graph of size n with maximum node degree d that contains no K_t-minor. Then, there exists $\lambda \in \mathbb{R}$ such that G is $\varphi(\mu)$-hyperfinite, where $\varphi(\mu) \leq \gamma^2[(3+\sqrt{6})d/\mu]^2$ and $\gamma = \min(\lambda t, t^{3/2})$.*

Theorems 2.13 and 2.14 can be deduced from Theorem 2 in [40]. Proposition 4.1 in [6] yields a generalization of this result to all minor-free graphs. A derivation of this generalized result is, e.g., conducted in the proof of Corollary 2 in [37]. In Section 2.2, we give proofs for Theorems 2.13 to 2.15 directly using separator theorems for trees [15], planar graphs [39] and minor-free graphs [5, 35] to obtain estimates of $\varphi(\mu)$ without hidden constants.

2.1.5. Carathéodory's Theorem

The convex hull of a set of points P is the minimal convex set that contains P.

Definition 2.16 (Convex hull). Let $P := \{\vec{\boldsymbol{p}}_1, \dots, \vec{\boldsymbol{p}}_\ell\} \subset \mathbb{R}^N$ be a finite set of points. The convex hull of P is defined as

$$\operatorname{conv}(P) := \left\{ \sum_{i=1}^{\ell} a_i \vec{\boldsymbol{p}}_i \;\middle|\; \forall i : a_i \in [0,1] \wedge \sum_{i=1}^{\ell} a_i = 1 \right\}. \qquad \blacksquare$$

The proof of the main result will make use of Carathéodory's Theorem, which is stated in Theorem 2.17 below. Recall that, given a finite set of points $P \subset \mathbb{R}^N$, each point p in the convex hull of P can be expressed as convex combination of P. An interesting question is the following: Is there an upper bound (better than $|P|$) on the number of points from P one has to use to express p as convex combination of these points? The question has been answered by Carathéodory [13] for finite and compact subsets of $\mathbb{R}^N$. His result was later generalized to arbitrary finite subsets of $\mathbb{R}^N$ by Steinitz [47]. It states that, for every finite subset $X \subset \mathbb{R}^N$, a point in the convex hull of X can be expressed as a convex combination of only $N+1$ points from X.

Theorem 2.17 (Carathéodory's theorem [13, 47]). *Let* $X := \{\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_\ell\} \subset \mathbb{R}^N$ *be a set of points. For every* $\vec{\boldsymbol{y}} \in \operatorname{conv}(X)$ *there exists a subset* $X' \subseteq X$ *such that* $|X'| \leq N+1$ *and* $\vec{\boldsymbol{y}} \in \operatorname{conv}(X')$*.*

2.2. Hyperfinite Graphs: Proofs of Theorems 2.13 to 2.15

This section covers proofs for the known facts that forests, planar graphs and minor-free graphs are hyperfinite (see Theorems 2.13 to 2.15). In the proofs, separator theorems are utilized to split graphs into components of constant size and the components' size is bounded as a function of the number of separator nodes.

In general, there is no proper subset $C \subsetneq V$ that partitions a graph into three sets $A, B, C \neq \emptyset$ such that no edge joins a node in A with a node in B. For example, consider a clique, where each pair of nodes is connected by an edge. There is no choice of $C \neq V$ that disjoins two sets A and B. However, for trees, planar graphs and minor-free graphs there exist separator theorems that state that there always is a small set of nodes that roughly cuts the graph in half.

The aforementioned separator theorems are discussed in Section 2.2.1. Some definitions are given in Section 2.2.2. The actual proofs for trees, planar graphs and minor-free graphs are conducted in Sections 2.2.3 to 2.2.5.

2.2.1. Separator Theorems

Let $G = (V, E)$ be a graph and let $u, v \in V$ be two nodes. If there exists a node $w \in V$ such that all paths from u to v visit w, then we say that w separates u and v. We can extend this definition from a single node to a set of nodes.

Definition 2.18 (Separator set). Let $G = (V, E)$ be a graph and let $A, B, C \subset V$ be a partitioning of V. Then, C is called a separator set iff every path from a node in A to a node in B crosses at least one node from C. ∎

Given a tree G, one can always find a single node that partitions the tree into two sets such that no set is greater than $\frac{2}{3}|V(G)|$.

Theorem 2.19 (Tree separator [cf. 15]). *Let G be a tree of size n. Then, the nodes of G can be partitioned into two sets $A := \sigma_A(G)$, $B := \sigma_B(G)$ and a single separator node $c := \sigma_C(G)$ such that $|A|, |B| \leq \frac{2}{3}n$.*

It has been proved by Ungar [51] that there exists a separator set C for every planar graph such that $|A|, |B| \leq \frac{2}{3}n$ and that C is in $\mathcal{O}(\sqrt{n}\log n)$. The bound on the size of C was improved to $\mathcal{O}(\sqrt{n})$ by Lipton and Tarjan [39] later.

Theorem 2.20 (Planar separators [39]). *Let G be a planar graph of size n. Then, the nodes of G can be partitioned into two sets $A := \sigma_A(G)$, $B := \sigma_B(G)$ and a separator set $C := \sigma_C(G)$ such that $|A|, |B| \leq \frac{2}{3}n$ and such that C contains no more than $\gamma\sqrt{n}$ nodes, where $\gamma \leq 2\sqrt{2}$.*

Alon et al. [5, 6] have proved that every K_t-free graph G of size n has a separator set of size $t^{3/2}\sqrt{n}$ that splits G into two subgraphs of size at most $\frac{2}{3}n$. Kawarabayashi and Reed [35] improved the asymptotics of this result by showing that there exists a separator set of size $\mathcal{O}(t\sqrt{n})$ as conjectured by Alon et al. [5].

Theorem 2.21 (Minor-free separators [5, 35]). *Let t be an integer and be G be a graph of size n that contains no K_t-minor. Then, there exists $\lambda \in \mathbb{R}$ such that the nodes of G can be partitioned into two sets $A := \sigma_A(G)$, $B := \sigma_B(G)$ and a separator set $C := \sigma_C(G)$ such that $|A|, |B| \leq \frac{2}{3}n$ and C contains no more than $\gamma\sqrt{n} = \min(\lambda t, t^{3/2}) \cdot \sqrt{n}$ nodes.*

One can prove Theorems 2.19 to 2.21 by taking advantage of the same idea, i.e., employing separator theorems multiple times until all components have constant size. In fact, one can carry out the proof that minor-free graphs are hyperfinite only and conclude that forests and planar graphs are hyperfinite, too. We are interested in good explicit bounds on $\varphi(\mu)$, and therefore we derive estimates separately. However, this affects details of the proofs only, and we can still give the same outline for all three proofs.

Proof outline We decompose the graph G into components of constant size by applying the corresponding separator theorem multiple times and removing all edges that are incident to separator nodes. Recall that we want to prove the results for degree-bounded graphs: If the number of separator nodes is bounded, then the number of removed edges is bounded, too. We construct a splitting tree where each application of the separator theorem is represented by a node and its two children. The parent node represents the graph before applying the separator theorem, and the two child nodes correspond to the resulting sets A and B. The separator set C is kept separately. We assign a level to each node of the tree and bound the number of separator nodes per level. Summing over all levels gives the desired bound.

2.2.2. Splitting Trees

A splitting tree is a structure that describes the repeated application of one of the aforementioned separator theorems to a graph (see Theorems 2.19 to 2.21). We construct a splitting tree $T(G)$ of a graph $G = (V, E)$ by recursively applying the corresponding separator theorem to it. The root of $T(G)$ is G. If $|V| > \varphi$, then we use the separator theorem to obtain subgraphs $\sigma_A(G)$, $\sigma_B(G)$ and $\sigma_C(G)$. We add $T(\sigma_A(G))$ and $T(\sigma_B(G))$ to the (empty) set of children of $T(G)$, children(G), and recurse into their construction. If, on the other hand, $|V| \leq \varphi$, we do nothing. When T is fully constructed, we assign a level $\mathrm{lvl}(T(G'))$ to each node $T(G')$. Leaves are on level 0. An inner node's level is one greater than the maximum level of its two child nodes. All separator vertices that are used to split nodes of $T(G)$ are gathered in the set $S(G)$. See Fig. 2.2 on the following page for an example of a splitting tree.

In the following definition, we always use the separator theorem that fits best, i.e., Theorem 2.19 for trees and so on. Without loss of generality, we assume that there always exists a uniquely defined separator set such that the splitting tree is uniquely defined, too. We are allowed to do this because we make use of the properties that are guaranteed for every separator set by the respective separator theorem only.

Definition 2.22 (Splitting tree). Let G be a tree, planar graph or minor-free graph such that one of the Theorems 2.19 to 2.21 can be applied to it. Then, the splitting tree $T(G)$, its level $\mathrm{lvl}(T(G))$ and its set of separators $S(G)$ are defined as follows.

$$
\begin{aligned}
T(G) &:= (G, \mathrm{children}(G)) \\
\mathrm{children}(G) &:= \begin{cases} \{T(\sigma_A(G)), T(\sigma_B(G))\} & \text{if } |V(G)| > \varphi \\ \emptyset & \text{otherwise} \end{cases} \\
\mathrm{lvl}(T(G)) &:= \begin{cases} 1 + \max_{T(G') \in \mathrm{children}(G)} \mathrm{lvl}(T(G')) & \text{if } \mathrm{children}(G) \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \\
S(G) &:= \begin{cases} \sigma_C(G) \cup \bigcup_{T(G') \in \mathrm{children}(G)} S(G') & \text{if } |V(G)| > \varphi \\ \emptyset & \text{otherwise}\,. \end{cases}
\end{aligned}
$$

We write $T_i(G)$ to denote all nodes of $T(G)$ on level i and, similarly, $S_i(G)$ to denote all separator vertices in nodes of $T(G)$ that are used to split nodes of $T_i(G)$. ∎

In the remainder of Section 2.2, we circumvent unnecessarily complicated statements by identifying a node $T(G')$ with its subgraph G'. To avoid any ambiguity

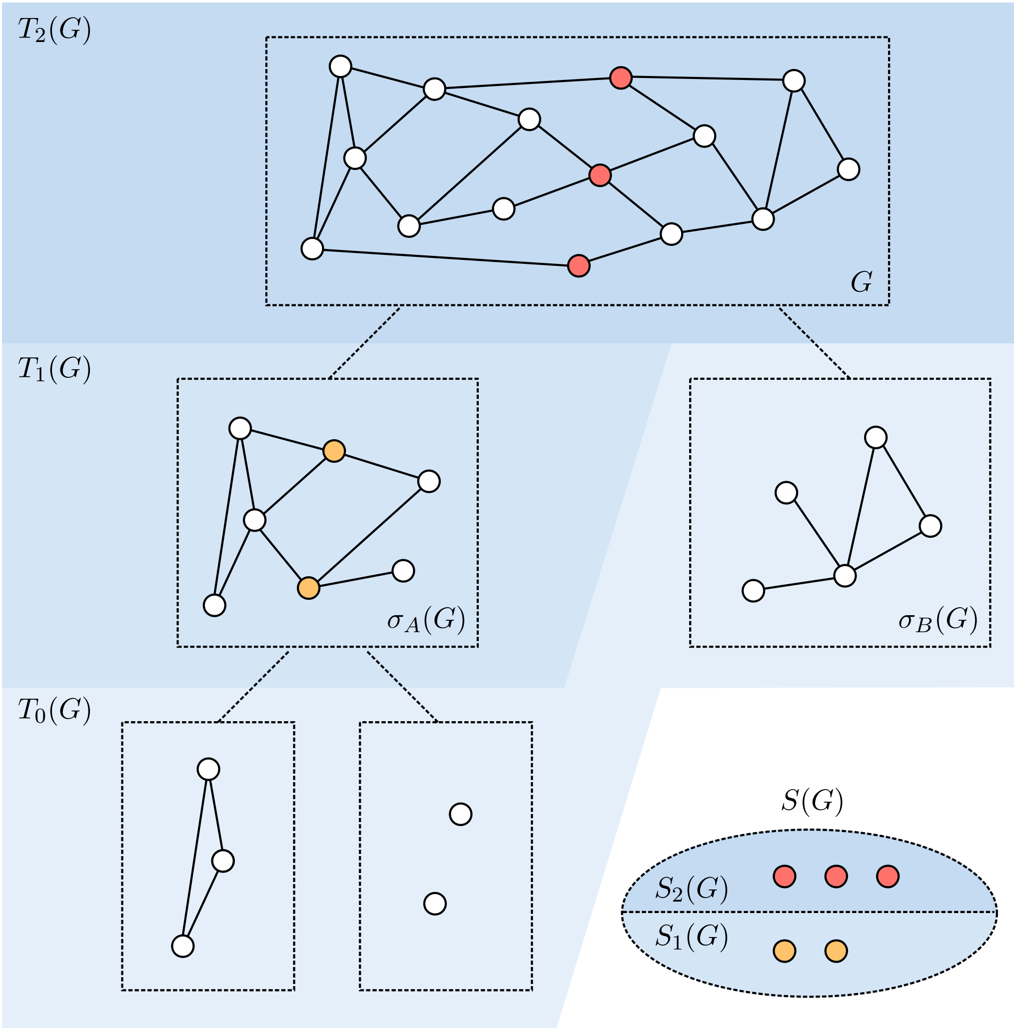


Figure 2.2.: A splitting tree of the graph depicted in Fig. 2.1 on page 17 with maximum component size $\varphi = 5$. Different levels are colored in different shades of blue. The separator vertices on the second level are colored red, and the separator vertices on the first level are colored orange.

that may arise, we refer to the elements of $T(G)$ as nodes and to the elements of $V(G)$ as vertices.

The following lemma states that the splitting tree provides the desired decomposition of G into connected components of size at most φ.

Lemma 2.23. *For every graph $G = (V, E)$ and its splitting tree $T(G)$, the graph obtained by joining the nodes in level zero and all separator vertices, i.e., $T_0(G)$ and $S(G)$, is a decomposition of G into components of size at most φ.*

Proof. The set $T_0(G) \cup S(G)$ is a decomposition of G because an arbitrary vertex $u \in V$ is an element of either exactly one of the leaf nodes in $T_0(G)$ or of $S(G)$, and therefore $V = \bigcup\{V(G') \mid (G', \cdot) \in T_0(G)\} \cup S(G)$. All components in nodes of $T_0(G)$ have size at most φ by construction, and $S(G)$ contains only isolated vertices. Since no edges are added, all components have a size of at most φ. □

The fact that the number of nodes on each level of the splitting tree and their accumulated size is bounded will be used in the proofs of Theorems 2.13 to 2.15.

Lemma 2.24. *Let G be a graph of size n and $T := T(G)$ be its splitting tree. At most $n/((\frac{3}{2})^{i-1}\varphi)$ nodes of T are on level i, and the total number of vertices in all nodes on level i is at most n.*

Proof. Let $\ell < \infty$ denote the maximum level of $T(G)$. Let $T(H)$ be an arbitrary node on level one and $T(H_p)$ its parent on level two. Since H was spit, it is clear that $|V(H)| > \varphi$. Its parent H_p was also split, and it holds by the respective separator theorem that

$$\frac{2}{3}|V(H_p)| \geq |V(H)| > \varphi \Rightarrow |V(H_p)| > \frac{3}{2}\varphi\,.$$

By induction, we conclude that an arbitrary node on level $i \geq 1$ has size at least $(3/2)^{i-1}\varphi$. An arbitrary vertex is contained in at most one of the nodes on level i as the level decreases on every path from the root to a leaf. Therefore, we can bound the total number of vertices in all components $H_1, \ldots, H_p$ on level i to at most n, i.e., $\sum_{j=1}^{p} |V(H_j)| \leq n$. Since every node on level i has size at least $(\frac{3}{2})^{i-1}\varphi$ and there are at most n vertices in the nodes on level i, there are at most $n/((\frac{3}{2})^{i-1}\varphi)$ nodes on level i of the splitting tree. □

2.2.3. Forests Are Hyperfinite

Theorem 2.13 (Forests are hyperfinite [cf. 40]—repeated). *Let $G = (V, E)$ be a forest of size n with maximum node degree d. Then, G is $\varphi(\mu)$-hyperfinite, where $\varphi(\mu) = 3d/\mu$.*

Proof. Let $\varphi \geq 1$ and μ be minimal such that G is (μ, φ)-hyperfinite. We construct a splitting tree according to Definition 2.22 by applying Theorem 2.19 and obtain a decomposition of G into components of size at most φ as stated by Lemma 2.23. Fix some $i \geq 1$ and let $T_i(G) = \{T(H_{i,1}), \ldots, T(H_{i,p_i})\}$ be the set of nodes on level i. We bound $|S_i(G)|$. Each node $H_{i,j}$ is split by a single vertex. As Lemma 2.24 states, there are at most $n/((\frac{3}{2})^{i-1}\varphi)$ nodes on level i of the splitting tree. Therefore, the number of separator vertices on level i is bounded by

$$|S_i(G)| \leq \sum_{j=1}^{p_i} 1 \leq \frac{n}{(\frac{3}{2})^{i-1}\varphi} = \frac{n}{\varphi}\left(\frac{2}{3}\right)^{i-1} .$$

The total number of separator vertices can be estimated by summing over all levels. As $\ell \to \infty$, the series converges to the limit of a geometric series.

$$|S(G)| = \sum_{i=1}^{\ell} |S_i(G)| \leq \frac{n}{\varphi} \sum_{i=1}^{\ell} \left(\frac{2}{3}\right)^{i-1} < \frac{n}{\varphi} \sum_{i=0}^{\infty} \left(\frac{2}{3}\right)^{i} = 3\frac{n}{\varphi} .$$

Each vertex in G is incident to at most d edges. An edge is removed iff it is incident to a vertex in $S(G)$, thus $d \cdot |S(G)| \geq \mu n$ by Definition 2.12. It follows that $d \cdot 3\frac{n}{\varphi} \geq \mu n$. Rearranging gives the desired result:

$$\mu n \leq d \cdot 3\frac{n}{\varphi} \Leftrightarrow \varphi \leq 3\frac{d}{\mu} . \qquad \square$$

2.2.4. Planar Graphs Are Hyperfinite

Theorem 2.14 (Planar graphs are hyperfinite [cf. 40]—repeated). *Let $G = (V, E)$ be a planar graph of size n with maximum node degree d. Then, G is $\varphi(\mu)$-hyperfinite, where $\varphi(\mu) = \gamma^2[(3+\sqrt{6})d/\mu]^2$ and $\gamma \leq 2\sqrt{2}$.*

Proof. Let $\varphi \geq 1$ and μ be minimal such that G is (μ, φ)-hyperfinite. We construct a splitting tree according to Definition 2.22 by applying Theorem 2.20 and obtain a decomposition of G into components of size at most φ as stated by Lemma 2.23. Fix some $i \geq 1$ and let $T_i(G) = \{T(H_{i,1}), \ldots, T(H_{i,p_i})\}$ be the set of nodes on level i. Each node $H_{i,j}$ is split using at most $\gamma\sqrt{|V(H_{i,j})|}$ vertices as separators by Theorem 2.20. We apply the Cauchy-Schwarz inequality to bound $|S_i(G)|$. Set $\vec{\mathbf{1}} = (1, \ldots, 1)$ and $\vec{\boldsymbol{s}} = \left(\gamma\sqrt{|V(H_{i,1})|}, \ldots, \gamma\sqrt{|V(H_{i,p_i})|}\right)$, where p_i denotes the number of nodes on level i. It holds that

$$|S_i(G)| \leq \sum_{j=1}^{p_i} \gamma\sqrt{|V(H_{i,j})|} = \left|\langle \vec{\mathbf{1}}, \vec{\boldsymbol{s}}\rangle\right| \leq ||\vec{\mathbf{1}}||_2 \cdot ||\vec{\boldsymbol{s}}||_2 = \sqrt{p_i} \cdot \gamma \sqrt{\sum_{j=1}^{p_i} |V(H_{i,j})|}\,.$$

As Lemma 2.24 states, there are at most $n/((\frac{3}{2})^{i-1}\varphi)$ nodes on level i of the splitting tree, i.e., $p_i \leq n/((\frac{3}{2})^{i-1}\varphi)$. The total number of vertices in all nodes on level i is at most n, i.e., $\sum_{j=1}^{p_i} |V(H_{i,j})| \leq n$. Therefore, we have that

$$|S_i(G)| \leq \ldots \leq \sqrt{p_i} \cdot \gamma \sqrt{\sum_{j=1}^{p_i} |V(H_{i,j})|} \leq \sqrt{\frac{n}{(\frac{3}{2})^{i-1}\varphi}} \cdot \gamma\sqrt{n} \leq n\frac{\gamma}{\sqrt{\varphi}} \left[\sqrt{\frac{2}{3}}\right]^{i-1} .$$

The total number of separator vertices $|S(G)|$ can be estimated by summing over all levels. As $\ell \to \infty$, the series converges to the limit of a geometric series.

$$|S(G)| = \sum_{i=1}^{\ell} |S_i(G)| \leq n\frac{\gamma}{\sqrt{\varphi}} \sum_{i=1}^{\ell} \left[\sqrt{\frac{2}{3}}\right]^{i-1} < n\frac{\gamma}{\sqrt{\varphi}} \sum_{i=0}^{\infty} \left[\sqrt{\frac{2}{3}}\right]^{i} = n\frac{\gamma}{\sqrt{\varphi}}(3+\sqrt{6})\,.$$

Each vertex in G is incident to at most d edges. An edge is removed iff it is incident to a vertex in $S(G)$, thus $d \cdot |S(G)| \geq \mu n$ by Definition 2.12. It follows that $d \cdot \frac{\gamma}{\sqrt{\varphi}}(3+\sqrt{6})n \geq \mu n$. Rearranging and plugging in $\gamma = 2\sqrt{2}$ from the planar separator Theorem 2.20 on page 22 gives the desired result:

$$\mu \leq d \cdot (3+\sqrt{6})\frac{\gamma}{\sqrt{\varphi}} \;\Leftrightarrow\; \sqrt{\varphi} \leq \gamma(3+\sqrt{6})\frac{d}{\mu} \;\Leftrightarrow\; \varphi \leq \gamma^2\left[(3+\sqrt{6})\frac{d}{\mu}\right]^2 . \qquad \square$$

2.2.5. Minor-Free Graphs Are Hyperfinite

Theorem 2.15 (Minor-free graphs are hyperfinite [cf. 6, 35]—repeated). *Let $G = (V, E)$ be a graph of size n with maximum node degree d that contains no K_t-minor. Then, there exists $\lambda \in \mathbb{R}$ such that G is $\varphi(\mu)$-hyperfinite, where $\varphi(\mu) \leq \gamma^2[(3+\sqrt{6})d/\mu]^2$ and $\gamma = \min(\lambda t, t^{3/2})$.*

Proof. This proof is conducted as the proof of Theorem 2.14 using Theorem 2.21 instead of Theorem 2.20, i.e., γ is substituted by $\min(\lambda t, t^{3/2})$. $\square$

3. Bound for Hyperfinite Graphs

In this chapter, we construct small graphs that share their local structure with a given degree-bounded hyperfinite graph of arbitrary size and derive upper bounds on their size. The central idea of this construction is to split the hyperfinite graph into pieces of constant size and prove that only a constant number of these components is required to construct the small graph. The result itself holds for every class of (μ, φ)-hyperfinite graphs, and the upper bound on the small graph's size depends on φ.

In Section 3.1, we formalize the problem, which has already been sketched in Section 1.1. Section 3.2 covers the main result and its proof. Results for trees, planar graphs and minor-free graphs, which state explicit upper bounds, are derived in Section 3.3. In Section 3.4, we encode a degree-bounded planar graph as a k-disk vector plus some edge insertions and removals and prove a lower bound on k.

3.1. Problem Formalization

The following problem was suggested by Alon and Lovász [33, Question 7]. Let $\mathcal{F}$ be a family of graphs. Recall that $M_{\mathcal{F}}(\epsilon, d, k)$ denotes the minimum size such that there exists, for every $G \in \mathcal{F}$, a graph H of at most this size and the ℓ_1 distance of the normalized k-disk vectors of G and H is at most ϵ (see Definition 2.6 on page 18). Alon has proved that $M_{\mathcal{F}}(\epsilon, d, k)$ attains a finite value for every family of graphs (including the family of all graphs; cf. proof recited by Lovász [42, Proposition 19.10]). The following formulation of this result and its proof are adapted to the context of this thesis.

Theorem 3.1 ([33, 42]). *Let $\mathcal{F}$ be the family of all graphs. For every $\epsilon > 0$ and all non-negative integers d and k, Definition 2.6 on page 18 is well-defined, i.e., $M_{\mathcal{F}}(\epsilon, d, k)$ is a positive integer.*

Proof. Set $N := N_{\mathcal{F}}(d, k)$ and denote the set $\mathcal{F}$ restricted to d-bounded graphs by $\mathcal{F}_d$. Let $\mathcal{F}_{\subseteq} = \{H_1, \dots, H_\ell\}$ be an inclusionwise maximal family of graphs in $\mathcal{F}_d$ such that $||\mathrm{freq}_k(H_i) - \mathrm{freq}_k(H_j)||_1 > \epsilon$ for all $1 \leq i < j \leq \ell$. The size of $\mathcal{F}_{\subseteq}$ is finite: all normalized k-disk vectors of graphs from $\mathcal{F}$ lie in $[0, 1]^N$, which has finite dimension and is bounded, and have pairwise ℓ_1 distance at least ϵ. For every pair of normalized k-disk vectors $\vec{\boldsymbol{u}}, \vec{\boldsymbol{v}} \in \mathcal{F}_{\subseteq}$, hyperspheres with radius $\epsilon/2$ around $\vec{\boldsymbol{u}}$ and $\vec{\boldsymbol{v}}$ do not intersect. Therefore, we have that $|\mathcal{F}_{\subseteq}| \leq 1/s$, where $s > 0$ is the volume of a hypersphere with radius $\epsilon/2$.

Given a graph $G \in \mathcal{F}_d$, denote the nearest graph in $\mathcal{F}_{\subseteq}$ by $f(G)$. Since $\mathcal{F}_{\subseteq}$ is maximal, there exists at least one graph $H_i \in \mathcal{F}_{\subseteq}$ such that $||\mathrm{freq}_k(G) - \mathrm{freq}_k(H_i)\,||_1 \leq \epsilon$ for every $G \in \mathcal{F}_d$. By definition, it holds that

$$\begin{aligned} M_{\mathcal{F}}(\epsilon, d, k) :=&\ \max_{G \in \mathcal{F}_d} \min \left\{ |G'| \;\middle|\; ||\mathrm{freq}_k(G) - \mathrm{freq}_k(G')\,||_1 \leq \epsilon \right\} \\ \leq&\ \max_{G \in \mathcal{F}_d} |f(G)| \leq \max_{H_i \in \mathcal{F}_{\subseteq}} |H_i| < \infty \,. \end{aligned} \qquad \square$$

The problem is to find an explicit estimate on $M(\epsilon) := M_{\mathcal{F}}(\epsilon, d, k)$. In this chapter, we give an explicit bound on $M(\epsilon)$ for the special case when a family of (μ, φ)-hyperfinite graphs is considered.

We note that the definition of $M(\cdot)$ in [33] uses the total variation distance $\delta(P, Q) = \sup_{A \subset \Omega} |P(A) - Q(A)|$ between two probability measures P and Q on a sample space Ω. For a finite number of elementary events (here: k-disks), this equals half the ℓ_1 distance $||P - Q||_1 = \sum_{x \in \Omega} |P(x) - Q(x)|$ [cf. 38, Section 4.1].

3.2. Main Result

In this section we show that, for any fixed non-negative integer k and every hyperfinite graph G with degree bounded by d, there exists a graph H with a similar k-disk vector whose size is independent of the size of G. Similarity is measured by $||\mathrm{freq}_k(G) - \mathrm{freq}_k(H)\,||_1$ and can be increased by allowing a larger, still constant size of H.

In the following, we assume that the upper bound on the degree of a graph is $d \geq 3$. The bounds can also be derived for $d < 3$. In fact, it is quite easy to construct small graphs with a similar k-disk vector for these graphs without the machinery used in the proof of Theorem 3.2. The cases $d = 0$ (only isolated nodes) and $d = 1$ (isolated nodes and single edges) are trivial: One has to combine isolated and paired nodes (i.e., edges) at the appropriate rate only. The case $d = 2$ (paths and cycles) has briefly been considered in Section 1.1. Here, one can join short paths and small cycles to obtain the desired graph.

The following theorem is the main result obtained in this chapter. Recall that $N_{\mathcal{F}}(d, k)$ denotes the total number of d-bounded k-disks of all graphs from a family of graphs $\mathcal{F}$ (see Definition 2.5 on page 17).

Theorem 3.2. *Let $\mathcal{F}$ be an arbitrary family of $\varphi(\mu)$-hyperfinite graphs, let $\epsilon \in (0,1]$ and $d \geq 3, k \geq 0$ be integers. Set $N := N_{\mathcal{F}}(d,k)$. Then, it holds that $N \leq 2^{\binom{3d^k/2}{2}}$ and that*

$$M_{\mathcal{F}}(\epsilon, d, k) \leq 12\frac{N}{\epsilon^2} \cdot \varphi\left(\frac{\epsilon}{9d^k}\right) .$$

In Section 3.3, we derive the following improved versions of the above theorem for forests, planar graphs and minor-free graphs by employing Theorems 2.13 to 2.15 on page 20.

Corollary 3.3. *Let $\mathcal{F}$ be the family of forests, $\epsilon \in (0,1]$ and $d \geq 3, k \geq 0$ be integers. Set $N := N_{\mathcal{F}}(d,k)$. Then, it holds that $N \leq d^k 2^{7.5d^k+2}$ and that*

$$M_{\mathcal{F}}(\epsilon, d, k) \leq 4 \cdot 10^2 \cdot \frac{d^{k+1}N}{\epsilon^3} = \mathcal{O}\left(\frac{d^{2k+1} \cdot 2^{7.5d^k}}{\epsilon^3}\right) .$$

Corollary 3.4. *Let $\mathcal{F}$ be the family of planar graphs, $\epsilon \in (0,1]$ and $d \geq 3, k \geq 0$ be integers. Set $N := N_{\mathcal{F}}(d,k)$. Then, it holds that $N \leq d^k 2^{7.5d^k+2}$ and that*

$$M_{\mathcal{F}}(\epsilon, d, k) \leq 3 \cdot 10^5 \cdot \frac{d^{2k+2}N}{\epsilon^4} = \mathcal{O}\left(\frac{d^{3k+2} \cdot 2^{7.5d^k}}{\epsilon^4}\right) .$$

Corollary 3.5. *Let $\mathcal{F}$ be the family of graphs with no K_t-minor, $\epsilon \in (0,1]$ and $d \geq 3, k \geq 0$ be integers. Set $N := N_{\mathcal{F}}(d,k)$. Then, it holds that $N \leq 2^{\binom{3d^k/2}{2}}$ and that there exists $\lambda \in \mathbb{R}$ such that*

$$M_{\mathcal{F}}(\epsilon, d, k) \leq \min\left(\lambda t^2, 3 \cdot 10^4 \cdot t^{9/4}\right) \cdot \frac{d^{2k+2}N}{\epsilon^4} = \mathcal{O}\left(\frac{t^2\, d^{2k+2}\, N}{\epsilon^4}\right) .$$

3.2.1. Outline of the Proof of Theorem 3.2

In Section 1.1, we cut out single connected components from cycles and grids to obtain small graphs that approximate their respective k-disk vectors. This idea can be generalized to hyperfinite graphs. Since the structure of hyperfinite graphs can be more complex compared to, e.g., cycle graphs, we do not hand-pick a single component of constant size.

Instead, a hyperfinite graph G is split into induced subgraphs $G[C_i]$ of size at most φ by Definition 2.12 on page 19. Removing edges alters all k-disks in G that originally contained a deleted edge. Therefore, the union of all $G[C_i]$ has a different k-disk vector than G. Nevertheless, the difference is small and we can express the normalized k-disk vector of G as a convex combination of the normalized k-disk vectors of all $G[C_i]$ plus the difference.

Next, we reduce the number of $G[C_i]$ that are involved in the construction. The number of coefficients greater than zero can be reduced to a constant number by using Carathéodory's theorem. We join multiple copies of the remaining $G[C_i]$ to reflect the size of their respective coefficients. It may be inevitable to make another error this way because the coefficients obtained from Carathéodory's theorem may be non-rational and the size of a component $G[C_i]$ may vary between 1 and φ.

However, the final graph H can be constructed by joining the appropriate number of copies of each component $G[C_i]$. It remains to sort out the formal details and estimate the error terms.

3.2.2. Proof of Theorem 3.2

Proof of Theorem 3.2. Let $G = (V, E)$ be a (μ, φ)-hyperfinite graph with maximum node degree d. Set $n := |V|$, $m := |E|$ and let $G[C_1], \ldots, G[C_\ell]$ denote the components of size at most φ obtained after removing at most μn edges according to Definition 2.12 on page 19. We use $G[C_1], \ldots, G[C_\ell]$ to construct a small graph with the proposed properties (see Fig. 3.1 on the following page).

First, we consider the graph G' obtained by joining $G[C_1], \ldots, [C_\ell]$, i.e., G without the edges between C_i and C_j for all $i \neq j$. As only edges are removed, it holds that $\bigcup_i C_i = V$. The normalized k-disk vector of G' is very similar to that of G. However, there is a difference, which is introduced by altered k-disks: Removing an edge $e = (u, v)$ alters all k-disks that contain this edge. Only k-disks with their root at distance at most k from both endpoints u and v can contain it. See Fig. 3.1 for a schematic visualization. We postpone the computation of a bound on this error and state its result only.

Lemma 3.6. *It holds that* $||\mathrm{freq}_k(G) - \mathrm{freq}_k(G')||_1 \leq 3\mu d^k$.

Next, we reduce the number of $G[C_i]$ that are involved in the construction. The normalized k-disk vector of G' is a convex combination of the normalized k-disk vectors of $G[C_1], \ldots, G[C_\ell]$. To see this, we rewrite it according to Definition 2.5 on

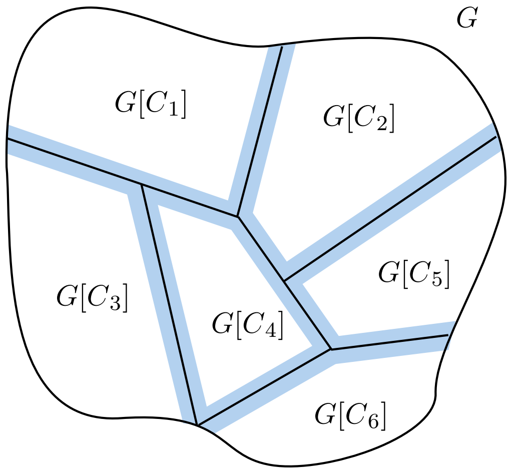


Figure 3.1.: Scheme of a hyperfinite graph's decomposition into six components $G[C_1], \ldots, G[C_6]$. nodes with changed k-disks are indicated in blue. The distance between such a node and a node of a removed edge is at most k.

page 17:

$$\begin{aligned}
\mathrm{freq}_k(G') = \frac{1}{n} \cdot \mathrm{dist}_k(G') &= \frac{1}{n} \sum_{i=1}^{\ell} \mathrm{dist}_k(G[C_i]) \\
&= \sum_{i=1}^{\ell} \frac{||\mathrm{dist}_k(G[C_i])\,||_1}{n} \frac{\mathrm{dist}_k(G[C_i])}{||\mathrm{dist}_k(G[C_i])\,||_1} \\
&= \sum_{i=1}^{\ell} \frac{|C_i|}{n} \cdot \mathrm{freq}_k(G[C_i]) \,. \qquad (3.1)
\end{aligned}$$

The ℓ_1 norm of a k-disk vector is equal to the number of k-disks covered by this vector. Also notice that the sum of all coefficients $a_i := |C_i|/n$ equals one, implying that (3.1) is a convex combination of normalized k-disk vectors.

We use Carathéodory's theorem (see Theorem 2.17 on page 21) to reduce the number of non-zero coefficients, i.e., the number of components involved in the convex combination. Let $b_1, \ldots, b_\ell$ denote the new coefficients obtained by applying Carathéodory's theorem. By Theorem 2.17 and Definition 2.5 on page 17, at most N of the b_i are non-zero. Without loss of generality, we assume that only $b_1, \ldots, b_N$ may be greater than zero, i.e., $b_{N+1} = \ldots = b_\ell = 0$. Then, it holds that

$$\mathrm{freq}_k(G') = \ldots = \sum_{i=1}^{\ell} a_i \cdot \mathrm{freq}_k(G[C_i]) = \sum_{i=1}^{N} b_i \cdot \mathrm{freq}_k(G[C_i]) \,. \qquad (3.2)$$

We aim at constructing a graph H similar to G (in terms of its normalized k-disk vector) by joining multiple copies of $G[C_1], \ldots, G[C_N]$. Suppose for a moment that (i) all b_i are rational and (ii) all components have equal size, i.e., $|C_1| = \ldots = |C_N| = \varphi$. Let b denote the lowest common denominator of $b_1, \ldots, b_N$. Joining $b_i \cdot b$ copies of $G[C_i]$ for every $i \in [N]$ would result in a (hypothetical) graph H'. This graph would have the normalized k-disk vector

$$\begin{aligned}
\mathrm{freq}_k(H') &\overset{(i)}{=} \frac{1}{|H'|} \sum_{i=1}^{N} b_i b \cdot \mathrm{dist}_k(G[C_i]) \\
&= \frac{1}{|H'|} \sum_{i=1}^{N} b_i b \cdot ||\mathrm{dist}_k(G[C_i])||_1 \cdot \frac{\mathrm{dist}_k(G[C_i])}{||\mathrm{dist}_k(G[C_i])||_1} \\
&\overset{(ii)}{=} \frac{1}{b\varphi} \sum_{i=1}^{N} b_i b \cdot \varphi \cdot \frac{\mathrm{dist}_k(G[C_i])}{||\mathrm{dist}_k(G[C_i])||_1} \\
&= \sum_{i=1}^{N} b_i \cdot \mathrm{freq}_k(G[C_i]) = \mathrm{freq}_k(G') \ .
\end{aligned}$$

It would hold that $||\mathrm{freq}_k(G) - \mathrm{freq}_k(H')||_1 \leq 3\mu d^k/2$ by Lemma 3.6 and Eq. (3.2). Unfortunately, some b_i may be non-rational, as well as some C_i may contain less than φ nodes. However, we will prove later that by slightly altering $b_1, \ldots, b_N$ and joining more than b copies of $G[C_1], \ldots, G[C_N]$ the following result can be achieved.

Lemma 3.7. *Let $c \geq 1$ and $\vartheta \geq 2$ be arbitrary integers. Then, there exist rational numbers $c_1, \ldots, c_N \in [0, 1]$ such that $\sum_{i=1}^{N} c_i = 1$ and the graph H constructed by joining $c_i c \lfloor \vartheta\varphi/|C_i| \rfloor$ copies of $G[C_i]$ for every $i \in [N]$ satisfies*

$$||\mathrm{freq}_k(G') - \mathrm{freq}_k(H)||_1 \leq \frac{N}{c} + \frac{1}{\vartheta - 1} \, .$$

By combining Lemma 3.6 and Lemma 3.7 we get the following result:

$$||\mathrm{freq}_k(G) - \mathrm{freq}_k(H)||_1 \leq 3\mu d^k + \frac{N}{c} + \frac{1}{\vartheta - 1} =: \epsilon \, .$$

Bounds on the parameter variables μ, c and ϑ as a function of the intended maximum error ϵ can be obtained by setting each summand to $\epsilon/3$ and rearranging them independently:

$$\begin{aligned}
\frac{\epsilon}{3} &= 3\mu d^k & \frac{\epsilon}{3} &= \frac{N}{c} & \frac{\epsilon}{3} &= \frac{1}{\vartheta - 1} \\
\Leftrightarrow \mu &= \frac{\epsilon}{9d^k} & \Leftrightarrow c &= 3\frac{N}{\epsilon} & \Rightarrow \vartheta &\leq \frac{4}{\epsilon} \, .
\end{aligned} \tag{3.3}$$

Then, $M(\epsilon)$ is bounded by

$$M(\epsilon) \le |V(H)| \le \sum_{i=1}^{N} c_i c \left\lfloor \vartheta \frac{\varphi}{|C_i|} \right\rfloor \cdot |C_i| \le c\vartheta\varphi \sum_{i=1}^{N} c_i = c\vartheta\varphi \le 12 \frac{N}{\epsilon^2} \varphi \,.$$

It remains to bound N. The size of a d-bounded k-disk is at most $3d^k/2$ by Observation 3.8 on the facing page. Let $\mathcal{F}$ be the family of all graphs and define $b := \binom{3d^k/2}{2}$. We construct an injective mapping $f : R_{\mathcal{F}}(d,k) \to \{0,1\}^b$ to bound $N = |R_{\mathcal{F}}(d,k)|$ by $|\{0,1\}^b| = 2^b$. Given a k-disk $D \in R_{\mathcal{F}}(d,k)$, we assign a unique number $g_D(u) \in \{0,\dots,3d^k/2\}$ to each node $u \in D$ such that the root of D maps to zero. Let h be a bijective function that maps every element of the set $\{(x,y) \mid x,y \in \{0,\dots,3d^k/2\} \wedge x < y\}$ to an element of $\{0,\dots,b\}$. We define $f(D) := (x_1,\dots,x_b)$ where $x_i = 1$ if $g_D^{-1}(h^{-1}(i)) \in E(D)$ and $x_i = 0$ otherwise.

Assume that f is not injective. Let $D_1, D_2 \in R_{\mathcal{F}}(d,k)$ such that $D_1 \neq D_2$ and $f(D_1) = f(D_2)$. Then, $g_{D_2}^{-1} \circ g_{D_1}$ is an isomorphism between D_1 and D_2. This yields a contraction because D_1 and D_2 are representatives of different equivalence classes. Therefore, it follows that $N = |R_{\mathcal{F}}(d,k)| \le |\{0,1\}^b| = 2^{\binom{3d^k/2}{2}}$. □

3.2.3. Deferred Proofs

It remains to prove Lemmas 3.6 and 3.7.

Proof of Lemma 3.6

Recall that $G[C_1],\dots,G[C_\ell]$ are obtained by removing edges according to Definition 2.12 on page 19 (see Fig. 3.1 on page 32). Deleting these edges alters all k-disks that originally contained such an edge. We bound the k-disk vectors' difference of G and the graph G' obtained by joining $G[C_1],\dots,G[C_\ell]$. First, we count all k-disks that contain a fixed edge. Then, we bound the total error by summing over all removed edges.

Lemma 3.6 (repeated). *It holds that* $||\mathrm{freq}_k(G) - \mathrm{freq}_k(G')||_1 \le 3\mu d^k$.

Proof. First, we bound the number of k-disks that are altered by removing a single edge. Removing such an edge $e = (u,v)$ alters all k-disks that contain this edge. A k-disk centered at a node w contains e iff there exists a path from u to w and a path from v to w, both of length at most k. We establish an upper bound on the number of possible w by counting only paths from u to w: There is only one path of length 0 (when $u = w$). Since the degree of a node is at most d, there are at most d possible extensions of this path. From each of these d nodes there are $d-1$ possible extensions (paths with ties are already covered by simple paths), and therefore at most $d(d-1)$ simple paths of length 2 exist. Induction gives the following upper bound on the number of altered k-disks:

$$1 + d + d(d-1) + \ldots + d(d-1)^{k-1} \leq \sum_{i=0}^{k} d^i = \frac{1 - d^{k+1}}{1-d} \leq \frac{d^{k+1}}{\frac{2}{3}d} = \frac{3d^k}{2}\,.$$

Actually, this is also an upper bound on the number of nodes that can be reached from u by a path of length at most k. Therefore, it also bounds the maximum size of a k-disk.

Observation 3.8. *The size of a d-bounded k-disk is at most $3d^k/2$.*

It remains to bound the error by summing over all μn removed edges. As each alternation of a k-disk results in one component of the k-disk vector being increased by one and another component being decreased by one, it holds that

$$||\text{freq}_k(G) - \text{freq}_k(G')||_1 \leq \frac{1}{n} \cdot \left(2\mu n \cdot \frac{3d^k}{2}\right) = 3\mu d^k\,. \tag{3.4}$$

□

Proof of Lemma 3.7

Recall the idea of combining copies of $G[C_i]$ in the proof of Theorem 3.2: We assumed that the coefficients $b_1, \ldots, b_N \in \mathbb{R}$ are rational and that all C_i are of equal size. Thus, it would have been possible to construct a small graph H' such that $\text{freq}_k(H') = \sum_{i=1}^{N} b_i \cdot \text{freq}_k(G[C_i]) = \text{freq}_k(G')$. However, both assumptions do not hold in general.

We make two modifications to the construction of H' to overcome this problem. First, we define new, rational coefficients $c_1, \ldots, c_N \in \mathbb{Q}$ by simply rounding $b_1, \ldots, b_N$. Second, we construct new graphs $G[D_1], \ldots, G[D_N]$ of almost equal size, which share their normalized k-disk vectors with $G[C_1], \ldots, G[C_N]$. This is done by joining $\lfloor \vartheta\varphi/|C_i| \rfloor$ copies of $G[C_i]$ for some integer $\vartheta \geq 2$. Using $c_1, \ldots, c_N$ and $G[D_1], \ldots, G[D_N]$, a graph H with the following properties can be constructed.

Lemma 3.7 (repeated). *Let $c \geq 1$ and $\vartheta \geq 2$ be arbitrary integers. Then, there exist rational numbers $c_1, \ldots, c_N \in [0,1]$ such that $\sum_{i=1}^{N} c_i = 1$ and the graph H constructed by joining $c_i c \lfloor \vartheta\varphi/|C_i| \rfloor$ copies of $G[C_i]$ for every $i \in [N]$ satisfies*

$$||\text{freq}_k(G') - \text{freq}_k(H)||_1 \leq \frac{N}{c} + \frac{1}{\vartheta - 1}\,.$$

Proof. We construct rational coefficients $c_1, \ldots, c_N \in \mathbb{Q}$. In particular, it must hold that $\sum_{i=1}^{N} c_i = 1$. We define two functions $f, g : [0,1] \to \mathbb{Q} \cap [0,1]$ as

$$f(b_i) := \max\left\{\frac{\nu}{c} \;\middle|\; \nu \in \{0, \ldots, c\} \wedge \frac{\nu}{c} \leq b_i\right\} = \frac{\lfloor b_i \cdot c \rfloor}{c}$$
$$g(b_i) := \min\left\{\frac{\nu}{c} \;\middle|\; \nu \in \{0, \ldots, c\} \wedge \frac{\nu}{c} > b_i\right\} = \frac{\lceil b_i \cdot c \rceil}{c}\,.$$

These functions map real coefficients to the nearest smaller or greater *allowed* (rational) coefficient, respectively. Since $f(b_i) \leq b_i$ and $g(b_i) > b_i$, it follows that

$$\sum_{i=1}^{N} f(b_i) \leq \sum_{i=1}^{N} b_i = 1 \leq \sum_{i=1}^{N} g(b_i)\,.$$

The new coefficients are defined as follows. Initialize all c_i by setting them to $f(b_i)$. If $\sum_{i=1}^{N} c_i = 1$, then we are finished. Otherwise, we increment b_1 by $\frac{1}{c}$, changing its value from $f(b_i) = \frac{\nu}{c}$ to $g(b_i) = \frac{\nu+1}{c}$. If $\sum_{i=1}^{N} c_i = 1$, then we are finished. Otherwise, we proceed to increment $c_2, c_3, \ldots$ until $\sum_{i=1}^{N} c_i = 1$. Since $1 \leq \sum_{i=1}^{N} g(b_i)$, the sum of all c_i will always equal one at some point.

Observation 3.9. *For every $i \in [N]$, it holds that $|b_i - c_i| \leq 1/c$. Additionally, it holds that $\sum_{j=1}^{N} c_j = 1$.*

The small graph H is constructed by joining $c_i c \cdot \lfloor \vartheta\varphi \,/\, |C_i| \rfloor$ copies of $G[C_i]$ for each $i \in [N]$. Let $G[D_i]$ denote the graph obtained by joining $\lfloor \vartheta\varphi \,/\, |C_i| \rfloor$ copies of $G[C_i]$. This way, we can also construct H by joining $c_i c$ copies of $G[D_i]$ for all $i \in \mathbb{N}$. Constructing $G[D_1], \ldots, G[D_N]$ pursues the idea of leveling the size of all $G[C_i]$. The normalized k-disk vectors of $G[C_i]$ and $G[D_i]$ are equal for all $i \in [N]$:

$$\mathrm{freq}_k(G[D_i]) = \frac{\mathrm{dist}_k(G[D_i])}{||\mathrm{dist}_k(G[D_i])\,||_1} = \frac{\lfloor \vartheta\varphi/|C_i| \rfloor \cdot \mathrm{dist}_k(G[C_i])}{\lfloor \vartheta\varphi/|C_i| \rfloor \cdot ||\mathrm{dist}_k(G[C_i])\,||_1} = \mathrm{freq}_k(G[C_i])\;.$$

Now, we show that the size of $G[D_i]$ is at least $(\vartheta - 1)\varphi$ and at most $\vartheta\varphi$ for every $i \in [N]$. It is easy to see that the upper bound holds:

$$|D_i| = \left\lfloor \frac{\vartheta\varphi}{|C_i|} \right\rfloor |C_i| \leq \frac{\vartheta\varphi}{|C_i|}|C_i| = \vartheta\varphi\,. \tag{3.5}$$

We can establish a lower bound on $|D_i|$ by exploiting the fact that each component $G[C_i]$ is of size at least one and of size at most $\varphi \geq 1$. Recall that $\vartheta \geq 2$.

$$|D_i| = \left\lfloor \frac{\vartheta\varphi}{|C_i|} \right\rfloor |C_i| \geq \frac{\vartheta\varphi}{|C_i|}|C_i| - 1 = \vartheta\varphi - \frac{|C_i|}{|C_i|} \geq \vartheta\varphi - \frac{\varphi}{1} = (\vartheta - 1)\varphi\,. \tag{3.6}$$

The normalized k-disk vector of H can be decomposed into a weighted sum of the normalized k-disk vectors of $G[C_i]$ in a way that is similar to the decomposition of the k-disk vector of G' in Eq. (3.2) on page 32. For every $i \in [N]$, we define the helper variable s_i:

$$s_i := c \cdot \frac{||\mathrm{dist}_k(G[D_i])\,||_1}{||\mathrm{dist}_k(H)\,||_1} - 1\,.$$

Now, the normalized k-disk vector of H can be decomposed into two sums:

$$\begin{aligned}
\mathrm{freq}_k(H) &= \frac{\mathrm{dist}_k(H)}{||\mathrm{dist}_k(H)\,||_1} \\
&= \frac{1}{||\mathrm{dist}_k(H)\,||_1} \sum_{i=1}^{N} c_i c \cdot \mathrm{dist}_k(G[D_i]) \\
&= \frac{1}{||\mathrm{dist}_k(H)\,||_1} \sum_{i=1}^{N} c_i c \cdot ||\mathrm{dist}_k(G[D_i])\,||_1 \frac{\mathrm{dist}_k(G[D_i])}{||\mathrm{dist}_k(G[D_i])\,||_1} \\
&= \sum_{i=1}^{N} \frac{c \cdot ||\mathrm{dist}_k(G[D_i])\,||_1}{||\mathrm{dist}_k(H)\,||_1} \cdot c_i \cdot \mathrm{freq}_k(G[D_i]) \\
&= \sum_{i=1}^{N} (1 + s_i) \cdot c_i \cdot \mathrm{freq}_k(G[D_i]) \\
&= \sum_{i=1}^{N} c_i \cdot \mathrm{freq}_k(G[C_i]) + \sum_{i=1}^{N} s_i c_i \cdot \mathrm{freq}_k(G[C_i])\,.
\end{aligned} \tag{3.7}$$

In the following, we bound s_i by $-\frac{1}{\vartheta-1} \le s_i \le \frac{1}{\vartheta-1}$ for every $i \in [N]$, which implies $|s_i| \le \frac{1}{\vartheta-1}$. Since $\mathrm{dist}_k(H) = \sum_{i=1}^{N} c_i c \cdot \mathrm{dist}_k(G[D_i])$ by the construction of H and $\sum_{i=1}^{N} c_i = 1$ (see Observation 3.9), it holds that

$$c(\vartheta-1)\varphi = \sum_{i=1}^{N} c_i c(\vartheta-1)\varphi \le ||\mathrm{dist}_k(H)\,||_1 \le \sum_{i=1}^{N} c_i c\vartheta\varphi = c\vartheta\varphi\,. \tag{3.8}$$

The proposed bounds on s_i can now easily be obtained by applying Eqs. (3.5), (3.6) and (3.8), respectively:

$$s_i = \frac{c \cdot ||\mathrm{dist}_k(G[D_i])\,||_1}{||\mathrm{dist}_k(H)\,||_1} - 1 \le \frac{c\vartheta\varphi}{c(\vartheta-1)\varphi} - 1 = \frac{\vartheta}{\vartheta-1} - 1 = \frac{1}{\vartheta-1}\,. \tag{3.9}$$

$$s_i = \frac{c \cdot ||\mathrm{dist}_k(G[D_i])\,||_1}{||\mathrm{dist}_k(H)\,||_1} - 1 \ge \frac{c(\vartheta-1)\varphi}{c\vartheta\varphi} - 1 = \frac{\vartheta-1}{\vartheta} - 1 = -\frac{1}{\vartheta} \ge -\frac{1}{\vartheta-1}\,. \tag{3.10}$$

Observation 3.10. *For every $i \in [N]$, it holds that $|s_i| \le 1\,/\,(\vartheta-1)$.*

Recall from Eq. (3.2) on page 32 that we have $\text{freq}_k(G') = \sum_{i=1}^N b_i \cdot \text{freq}_k(G[C_i])$. Using Eqs. (3.2) and (3.7), we establish the proposed bound on $||\text{freq}_k(G') - \text{freq}_k(H)\,||_1$:

$$
\begin{aligned}
&||\text{freq}_k(G') - \text{freq}_k(H)||_1 \\
&= \left|\left|\sum_{i=1}^N b_i \cdot \text{freq}_k(G[C_i]) - \sum_{i=1}^N c_i \cdot \text{freq}_k(G[C_i]) - \sum_{i=1}^N s_i c_i \cdot \text{freq}_k(G[C_i])\right|\right|_1 \\
&\leq \sum_{i=1}^N |b_i - c_i| \cdot ||\text{freq}_k(G[C_i])||_1 + \sum_{i=1}^N |s_i| \cdot c_i \cdot ||\text{freq}_k(G[C_i])||_1 . \\
&\leq \sum_{i=1}^N \frac{1}{c} \cdot 1 + \sum_{i=1}^N \frac{1}{\vartheta - 1} \cdot c_i \cdot 1 = \frac{N}{c} + \frac{1}{\vartheta - 1} .
\end{aligned}
$$

As $\text{freq}_k(G[C_i])$ is a normalized k-disk vector, its ℓ_1 norm equals one. Furthermore, it holds that $|b_i - c_i| \leq 1/c$ and $\sum_{i=1}^N c_i = 1$ (see Observation 3.9) and $|s_i| \leq 1\,/\,(\vartheta - 1)$ (see Observation 3.10). $\square$

3.3. Applications: Proofs of Corollaries 3.3 to 3.5

In this section, we apply the main result, Theorem 3.2, to forests, planar graphs and minor-free graphs. First, we establish an upper bound on the number of non-isomorphic planar k-disks in Section 3.3.1. This result is used in Sections 3.3.2 and 3.3.3 to derive closed explicit estimates of the small graph's size for trees and planar graphs by applying Theorems 2.13 and 2.14 on page 20. In Section 3.3.4 we apply Theorem 2.15 to derive a version of Theorem 3.2 for minor-free graphs.

3.3.1. On the Number of Planar k-Disks

Since the bound on the small graph's size $M_{\mathcal{F}}(\epsilon, d, k)$ in Theorem 3.2 depends on the dimension $N_{\mathcal{F}}(d, k)$ of the k-disk vector, i.e., the number of pairwise non-isomorphic, d-bounded k-disks in $\mathcal{F}$, it is desirable to calculate a better estimate of $N_{\mathcal{F}}(d, k)$ for planar graphs than it is given by Theorem 3.2.

The number of degree-bounded planar k-disks is bounded by the number of planar graphs (on a fixed number of nodes). Upper bounds on the number of (unlabeled) planar graphs on n nodes were, e.g., obtained by Poulalhon and Schaeffer [45] and Tutte [49]. The following result is due to Bonichon et al. [10, 11].

Theorem 3.11 ([10, 11]). *The number of non-isomorphic connected planar graphs of size n is at most $2^{4.91n} \leq 2^{5n}$.*

Although we did not introduce labeled graphs (except for rooted graphs where a single node is labeled as root), we note that this theorem refers to the number of unlabeled planar graphs.

Theorem 3.12. *Let $\mathcal{F}$ be the family of planar graphs on at least three nodes and $d \geq 3, k \geq 0$ be integers. Then, the number of pairwise non-isomorphic, d-bounded, planar k-disks $N_{\mathcal{F}}(d,k)$ is at most*

$$N_{\mathcal{F}}(d,k) \leq d^k 2^{7.5d^k+2} .$$

Proof. If a k-disk is d-bounded, then its size is at most $3d^k/2$ by Observation 3.8. Theorem 3.11 states that the number of connected planar graphs on n nodes is at most 2^{5n}. Additionally, k-disks are rooted graphs. There are at most n ways to root a graph on n nodes.

Combining these results yields that the number of d-bounded, planar k-disks on exactly n nodes is at most $n \cdot 2^{5n} \leq (3d^k/2) \cdot 2^{7.5d^k}$ where $n \leq 3d^k/2$. We sum over all possible $n \in \{1, \dots, 3d^k/2\}$ to get a bound on the number of all d-bounded, planar k-disks:

$$\sum_{i=1}^{3d^k/2} i \cdot 2^{5i} \leq \frac{3d^k}{2} \cdot \sum_{i=0}^{7.5d^k} 2^i = \frac{3d^k}{2} \cdot \frac{2^{7.5d^k+1} - 1}{2-1} \leq d^k 2^{7.5d^k+2} .$$

□

3.3.2. Bound for Forests

Recall Theorem 2.13 on page 20: It states that the family of d-bounded forests is $\mathcal{O}(d/\mu)$-hyperfinite. This result enables us to state a more intuitive formulation of Theorem 3.2 on page 29 for forests: it allows us to obtain an explicit bound on $|H|$.

Corollary 3.3 (repeated). *Let $\mathcal{F}$ be the family of forests, $\epsilon \in (0,1]$ and $d \geq 3, k \geq 0$ be integers. Set $N := N_{\mathcal{F}}(d,k)$. Then, it holds that $N \leq d^k 2^{7.5d^k+2}$ and that*

$$M_{\mathcal{F}}(\epsilon, d, k) \leq 4 \cdot 10^2 \cdot \frac{d^{k+1} N}{\epsilon^3} = \mathcal{O}\left(\frac{d^{2k+1} \cdot 2^{7.5d^k}}{\epsilon^3}\right) .$$

Proof. We use equation (3.3) from the proof of Theorem 3.2:

$$\mu = \frac{\epsilon}{9d^k} .$$

Plugging μ into the bound on φ from Theorem 2.13 on page 20 gives

$$\varphi(\mu) = 3\frac{d}{\mu} = 27 \cdot \frac{d^{k+1}}{\epsilon} .$$

Applying Theorem 3.2 gives the bound

$$M(\epsilon) \leq 12 \cdot \frac{N}{\epsilon^2}\varphi(\mu) = 324 \cdot \frac{d^{k+1}N}{\epsilon^3} \leq 324 \cdot \frac{d^{k+1} \cdot d^k 2^{7.5d^k+2}}{\epsilon^3} .$$

The bound on N follows from Theorem 3.12. □

3.3.3. Bound for Planar Graphs

We can adapt the proof of Corollary 3.3 to planar graphs. Recall that the family of planar graphs with degree bounded by d is $\mathcal{O}(d^2/\mu^2)$-hyperfinite by Theorem 2.14 on page 20.

Corollary 3.4 (repeated). *Let $\mathcal{F}$ be the family of planar graphs, $\epsilon \in (0,1]$ and $d \geq 3, k \geq 0$ be integers. Set $N := N_{\mathcal{F}}(d,k)$. Then, it holds that $N \leq d^k 2^{7.5d^k+2}$ and that*

$$M_{\mathcal{F}}(\epsilon, d, k) \leq 3 \cdot 10^5 \cdot \frac{d^{2k+2}N}{\epsilon^4} = \mathcal{O}\left(\frac{d^{3k+2} \cdot 2^{7.5d^k}}{\epsilon^4}\right) .$$

Proof. We use equation (3.3) from the proof of Theorem 3.2:

$$\mu = \frac{\epsilon}{9d^k} .$$

Plugging μ into the bound on φ from Theorem 2.14 on page 20 gives

$$\varphi(\mu) = \gamma^2 \left[(3+\sqrt{6}) \cdot \frac{d}{\mu}\right]^2 = \gamma^2 \left[9(3+\sqrt{6}) \cdot \frac{d^{k+1}}{\epsilon}\right]^2 .$$

Applying Theorem 3.2 yields the bound

$$M(\epsilon) \leq 12 \cdot \frac{N}{\epsilon^2}\varphi(\mu) = 972\gamma^2(3+\sqrt{6})^2 \cdot \frac{d^{2k+2}N}{\epsilon^4} \leq 230\,924 \cdot \frac{d^{2k+2} \cdot d^k 2^{7.5d^k+2}}{\epsilon^4} .$$

The bound on N follows from Theorem 3.12, and $\gamma \leq 2\sqrt{2}$ follows from Theorem 2.20 on page 22. □

3.3.4. Bound for Minor-Free Graphs

We can use the proof of Corollary 3.4 to show a similar result for minor-free graphs. As stated by Theorem 2.15 on page 20, the family of graphs that does not contain K_t as a minor is $\mathcal{O}(td^2/\mu^2)$-hyperfinite. However, the size of H depends on the size of the minor. We provide two bounds on the size of H: one is asymptotically tighter, the other one has no constant hidden inside asymptotic notation.

Corollary 3.5 (repeated). *Let $\mathcal{F}$ be the family of graphs with no K_t-minor, $\epsilon \in (0,1]$ and $d \geq 3, k \geq 0$ be integers. Set $N := N_{\mathcal{F}}(d,k)$. Then, it holds that $N \leq 2^{\binom{3d^k/2}{2}}$ and that there exists $\lambda \in \mathbb{R}$ such that*

$$M_{\mathcal{F}}(\epsilon, d, k) \leq \min\left(\lambda t^2, 3 \cdot 10^4 \cdot t^{9/4}\right) \cdot \frac{d^{2k+2}N}{\epsilon^4} = \mathcal{O}\left(\frac{t^2\, d^{2k+2}\, N}{\epsilon^4}\right).$$

Proof. This proof is conducted as the proof of Corollary 3.4 using Theorem 2.15 on page 20 instead of Theorem 2.14, i.e., γ is substituted by $\min(\lambda' t, t^{3/2})$ where λ' is chosen according to Theorem 2.21. Define $\lambda := (\lambda')^2$. We have that

$$M(\epsilon) \leq 12 \cdot \frac{N}{\epsilon^2}\varphi(\mu) = 972\gamma^2(3+\sqrt{6})^2 \cdot \frac{d^{2k+2}N}{\epsilon^4} \leq \min(\lambda t^2, 28\,866 \cdot t^{9/4}) \cdot \frac{d^{2k+2}N}{\epsilon^4}. \quad \square$$

3.4. Encoding Planar Graphs as k-Disk Vectors

Newman and Sohler [44] have proved that two planar graphs that have similar k-disk vectors can be transformed into each other by changing only a constant fraction of their edges. In this section, we encode a graph by its local view and a fill-in graph to form the global view. In particular, we encode a degree-bounded planar graph by a k-disk vector and a small amount of edges that need to be changed to obtain G after some *prototype* was recovered from the k-disk vector. We prove a lower bound on k (see Theorem 3.17 for the precise statement):

Theorem. *Let n be sufficiently large, $\epsilon \in \mathcal{O}(1/\log n - 2^{10}/n)$, $d \geq 5$ and $k \in \mathbb{N}$. If the family of d-bounded planar graphs of size n can be encoded by a k-disk vector of a graph of size at most N and ϵn edges, then $k > \log_d[\log_2([1/(3\log_2 n) - \epsilon]n)/10]$.*

3.4.1. Preliminaries on Encoding

The following definitions provide the basis for an encoding-based formulation of the result. An encoding maps the elements of a set A to the elements of a set B such that the mapping is reversible.

Definition 3.13 (Encoding). A finite set A can be encoded by a finite set B if there exists an injective function $f : A \to B$. In particular, it must hold that $|B| \geq |A|$. ■

Natural numbers can be encoded as follows.

Definition 3.14 (Binary encoding). Let $n \in \mathbb{N}$ and $z \in \{0, \ldots, 2^n - 1\}$. The function $\text{encode}_n : \{0, \ldots, 2^n - 1\} \to \{0, 1\}^n$ maps z to its binary representation of length n. ■

Inserting or removing an edge into or from a graph can be formalized as an edge operation.

Definition 3.15 (Edge operation). Let $n \in \mathbb{N}$. An edge operation on a graph of size n is a tuple (e, op) where $e \in [n] \times [n]$ and $op \in \{insert, delete\}$. ■

The number of planar graphs of size n grows exponentially in n. Actually, even the number of degree-bounded planar graphs of size n is of order $2^{\mathcal{O}(n)}$ [49]. Therefore, the exponential upper bound on the number of planar graphs from Theorem 3.11 cannot be improved to something non-exponential even for degree-bounded planar graphs. In the following, let $\mathcal{P}_{d,n}$ denote the set of d-bounded planar graphs of size n. There are at least 2^n of such graphs.

Lemma 3.16. *Let $d \geq 5, n \geq 9$ and $\mathcal{P}_{d,n}$ be the family of d-bounded planar graphs. It holds that $|\mathcal{P}_{d,n}| \geq 2^n$, i.e., there are more than 2^n pairwise non-isomorphic, d-bounded, planar graphs.*

Proof. See Section 3.4.3. □

3.4.2. Encoding a Graph by Its Local View and Its Global View

Let $X_{1,d,k,n} := \{\text{freq}_k(G_1) \mid G_1 \in \mathcal{P}_{d,n} \wedge G_2 \in \mathcal{P}_{d,N} \wedge \text{freq}_k(G_1) = \text{freq}_k(G_2)\}$ be the set of vectors that are normalized k-disk vectors of a d-bounded planar graph of size n as well as of a planar graph of size N, $X_{2,n,\epsilon} := \{(o_1, \ldots, o_\ell) \mid \ell \leq \epsilon n \wedge \forall i : o_i \text{ is edge operation on } n \text{ nodes}\}$ be the set of all sequences of at most ϵn edge operations on a node set of size n, and $X_{d,k,n,\epsilon} := X_{1,d,k,n} \times X_{2,n,\epsilon}$ be the Cartesian product of $X_{1,d,k,n}$ and $X_{2,n,\epsilon}$.

Theorem 3.17. *Let $n \geq 4.8 \cdot 10^4$, $0 < \epsilon < 1/(3 \log_2 n) - 2^{10}/n$, $d \geq 5$ and $k \in \mathbb{N}$. If $\mathcal{P}_{d,n}$ can be encoded by $X_{d,k,n,\epsilon}$, i.e., a k-disk vector of a graph of size at most N and ϵn edge operations, then $k > \log_d[\log_2([1/(3 \log_2 n) - \epsilon]n)/10]$.*

Proof. Suppose $f : \mathcal{P}_{d,n} \to X_{d,k,n,\epsilon}$ is an injective function. We define

$$\begin{aligned}
k(n,\epsilon,d) &:= \log_d \left[\log_2 \left[\left(\frac{1}{3\log_2 n} - \epsilon\right) n\right] /10\right] \\
N &:= N_{\mathcal{F}}(d,k),\ N' := \lceil \log_2 N \rceil \\
n' &:= \lceil \log_2 n \rceil \\
\mu &:= N\lceil \log_2 N\rceil + \lfloor \epsilon n \rfloor (2\lceil \log_2 n \rceil + 1)\,.
\end{aligned}$$

Notice that $k(n,\epsilon,d) > 0$ if n is greater than $4.8 \cdot 10^4$ and $\epsilon < 1/(3\log_2 n) - 2^{10}/n$.

Each pair of a k-disk and a sequence of at most ϵn edge operations can be mapped to a bit string of length μ by the function $g : X_{d,k,n,\epsilon} \to \{0,1\}^\mu$ that is defined as follows.

$$\begin{aligned}
&g((x_1,\dots,x_N),((u_1,v_1,o_1),\dots,(u_\ell,v_\ell,o_\ell))) \\
&\qquad := \mathrm{encode}_{N'}(x_1) \circ \ldots \circ \mathrm{encode}_{N'}(x_N) \circ \mathrm{encode}_{n'}(u_1) \circ \mathrm{encode}_{n'}(v_1) \\
&\qquad\qquad \circ\, \mathrm{encode}_1(o_1) \circ \ldots \circ \mathrm{encode}_{n'}(u_\ell) \circ \mathrm{encode}_{n'}(v_\ell) \circ \mathrm{encode}_1(o_\ell) \\
&\text{where } \ell \le \lfloor \epsilon n \rfloor \le \epsilon n\,.
\end{aligned}$$

Recall that every element in the image of $\mathrm{encode}_\ell(\cdot)$ is a bit string of length exactly ℓ. Therefore, g is injective, and every graph $G \in \mathcal{P}_{d,n}$ can be mapped to a unique bit string of length

$$\begin{aligned}
\mu &= N\lceil \log_2 N\rceil + \lfloor \epsilon n \rfloor (2\lceil \log_2 n \rceil + 1) \\
&\le N(\log_2 N + 1) + \epsilon n (2\log_2 n + 3) \\
&\le 3 \cdot (N \log_2 N + \epsilon n \log_2 n)
\end{aligned}$$

by $g(f(G))$. On the other hand, if B is a set of bit strings of equal length such that B encodes $\mathcal{P}_{d,n}$, then the length of the bit strings is at least $\log_2 |\mathcal{P}_{d,n}|$. Otherwise, the size of B is at most the number of bit strings of length $\ell < \log_2 |\mathcal{P}_{d,n}|$, i.e., $|B| \le 2^\ell < |\mathcal{P}_{d,n}|$, which yields a contradiction to the assumption that g is an encoding, which is injective by Definition 3.13. By Lemma 3.16, we have that

$$3 \cdot (N \log_2 N + \epsilon n \log_2 n) \ge \log_2 |\mathcal{P}_{d,n}| > n\,. \tag{3.11}$$

Assume that $k \le k(n,\epsilon,d)$. Then, it holds that

$$
\begin{aligned}
&3 \cdot (N \log_2 N + \epsilon n \log_2 n) \\
&\le 3 \cdot \left[d^k 2^{7.5d^k+2} \log_2\left(d^k 2^{7.5d^k+2}\right) + \epsilon n \log_2 n\right] \\
&\le 3 \cdot \left[2^{\log_2 d^k} 2^{8.5d^k} \log_2\left(2^{\log_2 d^k} 2^{8.5d^k}\right) + \epsilon n \log_2 n\right] \\
&\le 3 \cdot \left[2^{0.6d^k} 2^{8.5d^k} \log_2\left(2^{0.6d^k} 2^{8.5d^k}\right) + \epsilon n \log_2 n\right] \\
&\le 3 \cdot \left[2^{10d^{k(n,\epsilon,d)}} \log_2\left(2^{10d^{k(n,\epsilon,d)}}\right) + \epsilon n \log_2 n\right] \\
&= 3 \cdot \left[\left(\frac{1}{3\log_2 n} - \epsilon\right) n \log_2\left[\left(\frac{1}{3\log_2 n} - \epsilon\right) n\right] + \epsilon n \log_2 n\right] \\
&\le 3 \cdot \left[\left(\frac{1}{3\log_2 n} - \epsilon\right) n \log_2 n + \epsilon n \log_2 n\right] = n\,.
\end{aligned}
$$

This yields a contradiction to Eq. (3.11). □

3.4.3. Number of Degree-Bounded Planar Graphs: Proof of Lemma 3.16

The number of degree-bounded planar graphs of size n is of order $2^{\mathcal{O}(n)}$ [49]. We prove that there exist more than 2^n 5-bounded planar graphs of size n for every $n \ge 9$ by bounding the size of the family $\mathcal{L}_n$, which is defined as follows. See Fig. 3.2 for an illustration of the construction scheme.

Definition 3.18. Let $n \ge 9$. Define $\ell := \lfloor (n-7)/2 \rfloor$ and $V' := \{u_0, \ldots, u_{\ell+2}, v_{\ell-1}, \ldots, v_{\ell+2}\}$. Let $f : [\ell] \to \{0,1\}$ and $g : [\ell] \to \{0,1,2\}$ be two functions and let $G_{f,g} = (V,E)$ denote the graph where $V = V'$ if n is odd and $V = V' \cup \{w\}$ if n is even and

$$
\begin{aligned}
E = \; & \{(u_i, u_{i+1}) \mid i \in \{0,\ldots,\ell+1\} && \cup \{(v_i, v_{i+1}) \mid i \in \{-1,\ldots,\ell+1\} \\
& \cup \{(u_1, v_1), (u_{\ell+2}, v_{\ell+2})\} && \cup \{(v_i, u_{i+1}) \mid i \in [\ell] \wedge f(i) = 1\} \\
& \cup \{(v_i, u_{i+2}) \mid i \in [\ell] \wedge g(i) = 1\} && \cup \{(v_{i+1}, u_{i+1}) \mid i \in [\ell] \wedge g(i) = 2\}\,.
\end{aligned}
$$

The family $\mathcal{L}_n$ is defined as $\mathcal{L}_n := \{G_{f,g} \mid f : [\ell] \to \{0,1\}, g : [\ell] \to \{0,1,2\}\}$ where two isomorphic graphs $G_{f,g} \cong G_{f',g'}$ are contained only once. ■

Lemma 3.16 (repeated). *Let $d \ge 5, n \ge 9$ and $\mathcal{P}_{d,n}$ be the family of d-bounded planar graphs. It holds that $|\mathcal{P}_{d,n}| \ge 2^n$, i.e., there are more than 2^n pairwise non-isomorphic, d-bounded, planar graphs.*

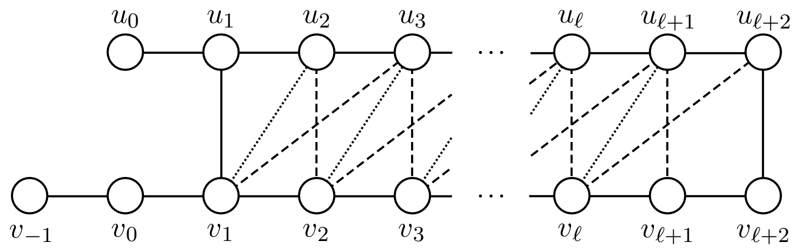


Figure 3.2.: Model of a graph $G \in \mathcal{L}_n$. Solid edges are present in every G. Dotted edges (v_i, v_{i+1}) may be absent. At most one dashed edge from each pair (v_i, u_{i+2}) and (v_i, u_i) may exist.

Proof. Without loss of generality assume that n is odd (the isolated node in graphs of even size from $\mathcal{L}_n$ does not affect the following reasoning) and define $\ell := \lfloor (n-7)/2 \rfloor$. Let $G_1 := G_{f_1,g_1} = (V, E_1)$ and $G_2 := G_{f_2,g_2} = (V, E_2)$ be two graphs from $\mathcal{L}_n$ where $V := \{u_0, \ldots, u_{\ell+2}, v_{\ell-1}, \ldots, v_{\ell+2}\}$. Although both graphs are defined on the same set of nodes, we refer to the nodes of G_2 as $x_i := u_i$ and $y_i := v_i$ to avoid any ambiguity.

First, we prove that there is at most one isomorphism function $h : V(G_1) \to V(G_2)$ if G_1 and G_2 are isomorphic. Therefore, G_1 and G_2 are isomorphic iff $f_1 \equiv f_2$ and $g_1 \equiv g_2$. We conclude the proof by showing that there are more than 2^n distinct pairs of functions f, g, which implies that $2^n < |\mathcal{L}_n| \leq |\mathcal{P}_{d,n}|$.

We prove that if $h : V(G) \to V(H)$ is an isomorphism function, then $h(u_i) = x_i$ and $h(v_i) = y_i$ for all i and j, respectively. Suppose that G_1 and G_2 are isomorphic and that h is an isomorphism function. There are exactly two nodes with degree one in each graph from $\mathcal{L}_n$, i.e., u_0 and v_{-1}. The node u_1 attached to u_0 has degree three, and the node v_0 attached to v_{-1} has degree two. Therefore, f must map the nodes as follows: $u_0 \to x_0$, $u_1 \to x_1$, $v_{-1} \to y_{-1}$ and $v_0 \to y_0$. Since there is only one node attached to v_0, it must also hold that $f(v_1) = y_1$ and $f(u_2) = x_2$.

We construct h incrementally and prove that it is defined uniquely. For this purpose, we place red tokens on u_2 and x_2 and blue tokens on v_1 and y_1. In turn i, we move the red token on G_1 along the edge (u_{i+1}, u_{i+2}) and the blue token on G_2 along (v_i, v_{i+1}). Then, we consider all possible movements of the tokens on G_2 along an edge such that h maps the nodes under tokens of the same color to each other while maintaining an isomorphism. If h is still unique after ℓ turns, we have proved that there is only one isomorphism function.

Consider turn i after moving the tokens on G_1, i.e., let $h(u_j) = x_j$ and $h(v_{j-1}) = y_{j-1}$ be determined for $j \leq i+1$ (see Fig. 3.3 on the following page). Regardless of the value of $f_2(i)$, we cannot move the blue token on G_2 to x_{i+1} because $h^{-1}(x_{i+1}) = u_{i+1}$

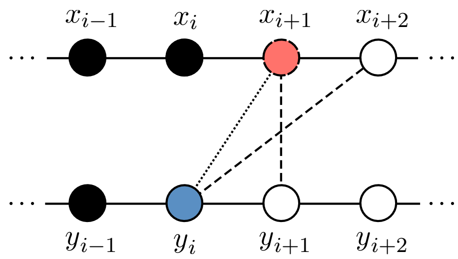


Figure 3.3.: In turn i, $h(u_j) = x_j$ and $h(v_{j-1}) = y_{j-1}$ for all $j \leq i+1$. Regardless of f_2 and g_2, the red token (dashed border) has to be moved to x_{i+2} and the blue token has to be moved to y_{i+1} in order to maintain h as an isomorphism function. Some edges are omitted.

is determined already. If $g_2(i) = 1$, we can move the blue token along (y_i, x_{i+2}). In this case, the edge (x_{i+1}, y_{i+1}) does not exist, which means that we cannot move the red token. If $g_2(i) = 2$ and we move the red token along (x_{i+1}, y_{i+1}), the same argument holds reversely, and we cannot move the blue token. Therefore, the only option to keep h an isomorphism function is to move the red token to x_{i+2} and the blue token to y_{i+1} such that $h(u_{i+2}) = x_{i+2}$ and $h(v_{i+1}) = y_{i+1}$.

The number of pairs (f, g) where $f : [\ell] \to \{0,1\}$ and $g : [\ell] \to \{0,1,2\}$ is $2^\ell \cdot 3^\ell$. Since two graphs G_{f_1,g_1} and G_{f_2,g_2} from $\mathcal{L}_n$ are isomorphic iff $f_1 \equiv f_2$ and $g_1 \equiv g_2$, the size of $\mathcal{L}_n$ is at least

$$|\mathcal{L}_n| = 2^\ell \cdot 3^\ell = 2^{\lfloor \frac{n-7}{2} \rfloor} \cdot 3^{\lfloor \frac{n-7}{2} \rfloor} \geq 2^{\frac{n-8}{2}} \cdot 2^{\log_2(3)\frac{n-8}{2}} > 2^n \,. \qquad \square$$

4. Further Results on k-Disk Vectors: Beyond Hyperfinite Graphs

Besides the results that were discussed in the preceding chapter, there are many more interesting aspects of k-disks and k-disk vectors. We discuss results regarding three of them in this chapter.

In Section 4.1, we examine expander graphs, which constitute a dual to hyperfinite graphs in a sense, and their k-disk vectors. In particular, we show that there exist *good* expanders and *well-separated* graphs that share the same k-disk vector (see Theorem 4.7). This result leads to some minor results on the feasibility of (normalized) k-disks, i.e., which vectors in $[0,1]^N$ appear as the normalized k-disk vector of a graph or in the limit of a graph sequence, which are discussed in Section 4.2. Moreover, we discuss the results of an empirical study of frequent k-disks from real world graphs like social networks and road networks in Section 4.3.

4.1. Expansion and k-Disk Vectors

In Chapter 3, we considered hyperfinite graphs, which are *weakly connected* from a global point of view because there is always a small fraction of their edges that can be removed to obtain many independent connected components. This section, however, deals with *expander graphs*, which are *well connected* graphs. The expansion of a graph G is high if all subsets of at most $|G|/2$ nodes have a large boundary relative to their size (cf. Definition 4.1 below). Expanders have many applications, e.g., in communication network design, where a message shall spread from a single source to a large number of participants quickly.

Understanding the relation between the *expansion* and the k-disk vector of a graph might be helpful to obtain a result like Theorem 3.2 for expander graphs, i.e., explicit upper bounds on the size of small graphs that have a normalized k-disk vector similar to that of an expander graph of arbitrary size. For example, one might conjecture that all graphs that *look like trees (with inner node degree at least three) locally* are good expanders.

In this section, we show that there exist, for sufficiently large (and even) n, good expanders as well as well-separated graphs of size n that have the same normalized k-disk vector. In other words, whether a graph is an expander or not is related to some information about its global structure, which is not covered by k-disks. Therefore, one can generally not conclude whether a graph is a good expander graph by looking at its k-disk vector only (see Theorem 4.7 for the precise statement):

Theorem. *There exists $n_0 \in \mathbb{R}$ such that for every even $n \geq n_0$ there exist two graphs of size n with the same k-disk vector: one with expansion at least $d/2 - \sqrt{d-2}$ and one with expansion at most $4/n$.*

On the contrary, triangle-freeness can for example be tested by analyzing all k-disks of a given graph. In the following, we give some definitions and results related to expander graphs first. Then, we employ these to obtain the aforementioned result.

4.1.1. Preliminaries on Expander Graphs

The following definition of edge expansion belongs to the so-called *combinatorial* characterizations of expansion. Loosely speaking, a graph is a good edge expander if each subset of at most half the graph's size has a large boundary.

Definition 4.1 (Edge expansion). The edge expansion of a graph $G = (V, E)$ is defined as

$$h(G) := \inf_{\{S | S \subset V \wedge 1 \leq |S| \leq |V|/2\}} \frac{|\partial S|}{|S|}\,. \qquad \blacksquare$$

A graph G is called a **c-expander (graph)** iff $h(G) \geq c$. We say that a sequence of graphs is an **c-expander (graph) sequence** iff there exists a constant $c > 0$ such that all graphs of the sequence are c-expanders.

Graphs that are not connected are not expanders because there is always a connected component of size at most $|V|/2$, and connected components have an empty boundary. Therefore, a graph G that is not connected has edge expansion $h(G) = 0$. However, this does not imply anything about the edge expansion of the connected components itself.

Another combinatorial definition of expansion is *node expansion* (where the number of edges on the boundary is replaced by the number of nodes on the outside of the boundary), which is related to edge expansion by a factor of d for d-bounded graphs. If the graph is d-regular, one can also give an algebraic characterization of expansion based on the eigenvalues of the graph's adjacency matrix. Refer to [23, 30] for surveys about expander graphs and more details.

We stick to edge expansion and point out the following relation instead. The edge expansion of a regular graph is related to the spectral gap of its adjacency matrix, i.e., the difference of the largest and the second largest eigenvalue. For any d-regular graph, it is known that the former one equals* d. The following theorem is a discrete version of a theorem by Cheeger [14] and Buser [12]. It has been proved by Dodziuk [17] and Alon and Milman [4] independently .

Theorem 4.2 (Cheeger inequalities [4, 17]). *Let G be a connected, d-regular graph and let λ be the second largest eigenvalue of its adjacency matrix. Then,*

$$\frac{d-\lambda}{2} \leq h(G) \leq \sqrt{2d(d-\lambda)}\,.$$

Random graphs are probability distributions over graphs. A well-known example is the uniform distribution over all graphs of size n. In the following definition, we consider the uniform distribution over all d-regular graphs of size n instead.

Definition 4.3 (Random d-regular graph). Let $\mathcal{F}_{n,d}$ be the family of all d-regular graphs of size n where n or d is even. The uniform distribution over $\mathcal{F}_{n,d}$ is denoted by $\mathcal{G}_{n,d}$, i.e., each $G \in \mathcal{F}_{n,d}$ is assigned the probability $1/|\mathcal{F}_{n,d}|$. ■

Alon [2] conjectured that, for large n, most random d-regular graphs are good expanders (in terms of a small eigenvalue gap). Eventually, this has been proved by Friedman [21].

Theorem 4.4 ([21]). *Let $\epsilon > 0$. Then, it holds with probability $1 - o(1)$ that a graph G drawn from $\mathcal{G}_{n,d}$ satisfies*

$$|\lambda_i| \leq 2\sqrt{d-1} + \epsilon$$

for all $2 \leq i \leq n$, where λ_i is the i-th largest eigenvalue of G.

Finite regular graphs contain cycles inevitably (if the degree is greater than one). On the other hand, Wormald [52] has proved that, as n goes to infinity, a constant fraction of all random regular graphs of size n contain no short cycles.

Theorem 4.5 ([52]). *The girth of a graph drawn from $\mathcal{G}_{n,d}$ is at least g with probability*

$$\exp\left(-\sum_{i=3}^{g-1} \frac{(d-1)^i}{2i} + o(1)\right)\,.$$

*Proof: Let A be the adjacency matrix of a d-regular graph G on n nodes. Since the row sums of A equal d, $A\vec{\mathbf{1}} = d\vec{\mathbf{1}}$ where $\vec{\mathbf{1}} = (1, \dots, 1)^T$. Therefore, d is an eigenvalue of A. Let $\vec{\boldsymbol{u}} = (u_1, \dots, u_n)$ be an eigenvector for the largest eigenvalue λ of A. Without loss of generality assume that at least one entry of $\vec{\boldsymbol{v}}$ is positive (otherwise consider $-\vec{\boldsymbol{v}}$). Let $j = \arg\max_i v_i$ be the index of the largest entry of $\vec{\boldsymbol{v}}$. Since $(\lambda, \vec{\boldsymbol{v}})$ is an eigenpair and the row sums of A equal d, it holds that $\lambda v_j = (A\vec{\boldsymbol{u}})_j \leq dv_j$. Therefore, d is also the largest eigenvalue of A.

The following k-disk is of special interest to us later.

Definition 4.6. The rooted tree where the distance from the root to each leaf is k and each inner node is d-regular is denoted by $\Delta_{d,k}$. ■

4.1.2. Expansion Information Is Lost in k-Disk Vectors

One might conjecture that graphs where all k-disks contain no cycles are good expanders because there are no edges that can be changed in order to increase the k-disk's boundary while keeping the k-disk connected. This intuition is supported by the fact that the infinite, d-regular tree T_d, which complies with the condition that all k-disks are trees, is a very good expander. Actually, it has edge expansion $d-2$ [cf. 30, Section 5.1]. This does not hold for finite trees, though. The separator theorem for trees (see Theorem 2.19 on page 21) implies that, for every finite tree T, there exists a subtree T' of size at least $|T|/3$ such that the boundary $\partial(V(T'))$ has size at most $d-1$, and therefore $h(T) \leq 3d/|T|$. Additionally, Alon [3] has proved that there exists a constant $c > 0$ such that, for every positive integer d and $n > 40d^9$, $h(G) \leq d/2 - c\sqrt{d}$ for every (finite) d-regular graph G on n nodes.

While every k-disk in T_d is isomorphic to $\Delta_{d,k}$, this is not the case for any finite tree, e.g., the k-disks of leaves are not isomorphic to $\Delta_{d,k}$. However, we will prove that there exists, for sufficiently large and even n, a good expander of size n as well as a well-separated graph of size n such that both have only k-disks that are isomorphic to $\Delta_{d,k}$.

Theorem 4.7. *For every $\epsilon > 0$, $d \geq 3$, and $k \geq 1$ there exist two sequences $(G_i)_{i\in\mathbb{N}}$ and $(H_i)_{i\in\mathbb{N}}$ such that $|G_i| = |H_i| =: n_i$, $n_i < n_{i+1}$, $\mathrm{freq}_k(G_i, \Delta_{d,k}) = \mathrm{freq}_k(H_i, \Delta_{d,k}) = 1$ but $h(G_i) \geq d/2 - \sqrt{d-1} - \epsilon$ whereas $h(H_{n_i}) \leq 4/n_i$ for every $i \in \mathbb{N}$.*

The proof of this theorem is based on the following two lemmas.

Lemma 4.8. *For every $\epsilon > 0$, $d \geq 3$ and $k \geq 1$, there exists an even integer $\nu := \nu(k) > 0$ and an expander sequence $(G_i)_{i\in\mathbb{N}}$ such that $|G_i| = \nu + 2i$, $\mathrm{freq}_k(G_i, \Delta_{d,k}) = 1$ and $h(G_i) \geq d/2 - \sqrt{d-1} - \epsilon > 0$ for every $i \in \mathbb{N}$.*

Lemma 4.9. *For every $d \geq 3$ and $k \geq 1$ there exists a sequence of graphs $(H_i)_{i\in\mathbb{N}}$ such that $|H_i| = 2\nu + 4i$ (where $\nu := \nu(k)$ is defined as in Lemma 4.8), $\mathrm{freq}_k(H_i, \Delta_{d,k}) = 1$ and $h(H_i) = 4/|H_i|$ for every $i \in \mathbb{N}$.*

Since most of the claims in Theorem 4.7 are covered by the preceding lemmas, the proof is a short deduction only. However, it remains to prove Lemmas 4.8 and 4.9.

Proof of Theorem 4.7. Let $(G_j)_{j\in\mathbb{N}}$ and $(H_i)_{i\in\mathbb{N}}$ be two sequences obtained from Lemmas 4.8 and 4.9, respectively. If there exists $j' \in \mathbb{N}$ for every $i \in \mathbb{N}$ such that $|G_{j'}| = |H_i|$, we are done because all other claims hold for $(G_j)_{j\in\mathbb{N}}$ and $(H_i)_{i\in\mathbb{N}}$ independently of each other.

Define $\nu := \nu(k)$ as in Lemma 4.8. We construct a sequence $(G_{j'})_{j'}$ by restricting $(G_j)_{j\in\mathbb{N}}$ to those G_j where $j = j'(i) := 2i + \nu/2$ for some $i \in \mathbb{N}$. To see that $|G_{j'(i)}| = |H_i|$ for all $i \in \mathbb{N}$, note that

$$|G_{j'(i)}| = \nu + 2j'(i) = \nu + 2(2i + \nu/2) = 2\nu + 4i = |H_i|$$

by Lemmas 4.8 and 4.9, respectively. □

Proof of Lemma 4.8

We prove the existence of $(G_i)_{i\in\mathbb{N}}$ from Lemma 4.8 by a probabilistic argument. The following equivalence serves as a link between $\Delta_{d,k}$ and the k-disks of random d-regular graphs (see Theorem 4.5).

Lemma 4.10. *Let $d \geq 3, k \geq 1$ and G be a d-bounded graph. Then, the girth of G is at least $2k+2$ iff for all d-bounded k-disks Δ that are not trees we have* $\operatorname{freq}_k(G,\Delta) = 0$.

Proof. Let G be a d-bounded graph with girth at least $2k+2$ and $r \in G$ be an arbitrary node. Since the distance between r and every node $u \in \operatorname{disk}_k(G,r)$ is at most k, the distance between to arbitrary nodes $u, v \in \operatorname{disk}_k(G,u)$ is at most $2k$. We claim that there exist no cycles in $\operatorname{disk}_k(G,r)$, i.e., $\operatorname{disk}_k(G,r)$ is a tree. This is certainly true for cycles of length less than $2k+2$ because the girth of G is at least $2k+2$. Suppose that there exists a cycle $C \subseteq \operatorname{disk}_k(G,r)$ of length greater than or equal to $2k+2$ in $\operatorname{disk}_k(G,r)$. Pick an arbitrary edge $(u,v) \in C$ (see Fig. 4.1a on the following page). The distances between r and u and r and v are at most k. The distance between u and v equals one. It follows that there exists a shorter cycle of length at most $2k+1$ in $\operatorname{disk}_k(G,r)$, which contradicts the fact that the girth of G is at least $2k+2$.

Now, let G be a d-bounded graph and $\operatorname{freq}_k(G,\Delta) = 0$ for all d-bounded k-disks Δ that are no trees. Assume that there exists a cycle C of length $\ell < 2k+2$ in G. Let $r \in C$ be an arbitrary node of this cycle. At least $\min(\ell, 2k)$ edges of C are contained in $\operatorname{disk}_k(G,r)$ as every path of length at most k that starts at r is contained in the k-disk (see Fig. 4.1b). In particular, all nodes of C are also nodes of the k-disk because $\ell \leq 2k+1$. Since a k-disk is an induced subgraph, $C \subseteq \operatorname{disk}_k(G,r)$, and therefore $\text{disk}_k(G, r)$ is not a tree. This contradicts the fact that all k-disks are trees. □

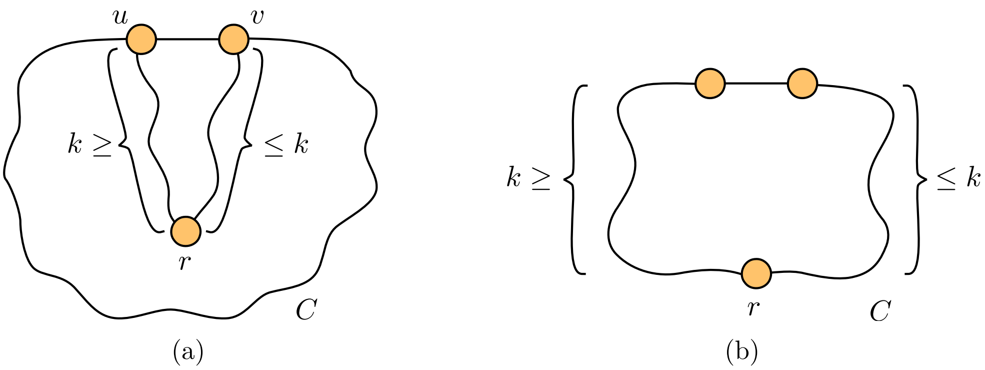


Figure 4.1.: (a) A cycle $C \subseteq \text{disk}_k(G, r)$ of length at least $2k + 2$ induces a shorter cycle of length at most $2k + 1$. (b) If all k-disks are trees, there cannot exist a cycle C of length at most $2k + 1$ because for every $r \in C$, $\text{disk}_k(G, r)$ would contain a cycle.

The following corollary of the preceding lemma is the actual result we use to prove Lemma 4.8. We restrict the preceding lemma to the case where the graph is not only d-bounded but also d-regular to obtain a stronger result.

Corollary 4.11. *Let $d \geq 3$ be an integer and G be a d-regular graph. Then, the girth of G is at least $2k + 2$ iff $\text{freq}_k(G, \Delta_{d,k}) = 1$.*

Proof. Let G be a d-regular graph with girth at least $2k+2$ and $u \in G$ be an arbitrary node. As shown in the proof of Lemma 4.10, $\text{disk}_k(G, r)$ is a tree. Since all nodes in G have degree d, nodes in $\text{disk}_k(G, r)$ that have a degree less than d can only exist because their neighborhood was not explored. Therefore, the distance between any leaf of $\text{disk}_k(G, r)$ and the root r equals k. On the other hand, the distance between any inner node of $\text{disk}_k(G, r)$ and the root r is at most $k - 1$, which means that all of their d neighbors are part of the k-disk, too. Summing up, $\text{disk}_k(G, r)$ is a (rooted) tree, all inner nodes have degree d and the distance between an arbitrary leaf and the root is k, i.e., $\text{disk}_k(G, r)$ is isomorphic to $\Delta_{d,k}$.

Note that $\text{freq}_k(G, \Delta_{d,k}) = 1$ implies $\text{freq}_k(G, \Delta) = 0$ for all d-bounded k-disks Δ that are not trees. Therefore, the other direction follows immediately from Lemma 4.10. □

Now, we use a probabilistic argument to prove Lemma 4.8, i.e., that there exists an expander sequence that contains, for every $n \in \mathbb{N}$, a graph G_n such that $|G_n| \geq n$ and $\text{freq}_k(G_n, \Delta_{d,k}) = 1$. Essentially, we combine Theorems 4.4 and 4.5 and use Corollary 4.11 to deduce the desired result.

Lemma 4.8 (repeated). *For every $\epsilon > 0$, $d \geq 3$ and $k \geq 1$, there exists an even integer $\nu := \nu(k) > 0$ and an expander sequence $(G_i)_{i \in \mathbb{N}}$ such that $|G_i| = \nu + 2i$, $\operatorname{freq}_k(G_i, \Delta_{d,k}) = 1$ and $h(G_i) \geq d/2 - \sqrt{d-1} - \epsilon > 0$ for every $i \in \mathbb{N}$.*

Proof. We show that, for every $k \geq 1$, there exists an even integer $\nu \geq 2$ such that, for every even $n \geq \nu$, there exists a d-regular graph on n nodes with girth at least $2k+2$. As Corollary 4.11 states, such a graph has the desired property. Let G be a graph drawn from $\mathcal{G}_{n,d}$. We define two events, E_{exp} and E_{girth}. Let E_{exp} denote the event that $|\lambda| \leq 2\sqrt{d-1} + 2\epsilon$, where λ is the second largest eigenvalue of G, and let E_{girth} denote the event that the girth of G is at least $2k+2$. If E_{exp} occurs, we have that

$$h(G) \geq \frac{d-\lambda}{2} \geq \frac{d - (2\sqrt{d-1} + 2\epsilon)}{2} \geq \frac{d}{2} - \sqrt{d-1} - \epsilon$$

by Theorem 4.2. If $\Pr(E_{\text{exp}} \cap E_{\text{girth}}) > 0$, then there exists at least one d-regular graph on n nodes that is an expander and has girth at least $2k+2$.

By Theorem 4.5, there exists an even integer ν_1 such that for all even $n \geq \nu_1$, the probability that a graph drawn from $\mathcal{G}_{n,d}$ has girth at least $2k+2$ is at least $p := \exp(-\frac{1}{2}\sum_{i=3}^{2k+1}(d-1)^i/i - 1) > 0$. On the other hand, it is likely that a large, random d-regular graph is an expander. More precisely, it is stated by Theorem 4.4 that there exists an even integer ν_2 such that for all even $n \geq \nu_2$, the probability that a graph drawn from $\mathcal{G}_{n,d}$ satisfies $|\lambda_i(G)| \leq 2\sqrt{d-1} + \epsilon$ is at least $1 - p/2$. We set $\nu := \max(\nu_1, \nu_2)$. Then, for every even $n \geq \nu$, it holds that

$$\begin{aligned}\Pr(E_{\text{exp}} \cap E_{\text{girth}}) &= \Pr(E_{\text{exp}}) + \Pr(E_{\text{girth}}) - \Pr(E_{\text{exp}} \cup E_{\text{girth}}) \\ &\geq (1 - p/2) + p - 1 \geq p/2 > 0\,.\end{aligned}$$

In combination with Corollary 4.11, this implies that, for every $n_i \in \{\nu + 2i \mid i \in \mathbb{N}\}$, there exists a d-regular graph G_i of size n_i that satisfies $\operatorname{freq}_k(G_i, \Delta_{d,k}) = 1$ and $h(G) \geq d/2 - \sqrt{d-1} - \epsilon$. □

Proof of Lemma 4.9

The expander graphs obtained from applying Lemma 4.8 can be used to construct graphs that have very small edge expansion. The main idea is to join such a graph with a copy of itself and cross an edge, i.e., swap one of its endpoints with the corresponding endpoint in the other copy. First, we describe the construction and prove its properties independently from the graphs obtained from Lemma 4.8.

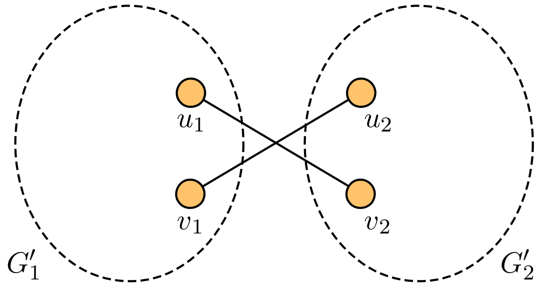


Figure 4.2.: The graph H. The nodes $V(H) \setminus \{u_1, u_2, v_1, v_2\}$ are omitted.

Lemma 4.12. *Let $d \geq 3, k \geq 1$ be integers and n be even. Let G be a connected d-bounded graph on $n/2$ nodes such that G is not a tree, but no k-disk contains a cycle. If G exists, then there exists a connected d-bounded graph H on n nodes such that* $\mathrm{freq}_k(G) = \mathrm{freq}_k(H)$ *and* $h(H) = 4/n$.

Proof. Suppose that G exists. Otherwise, there is nothing to show. By the assumption, there exists at least one cycle C in G. Pick an arbitrary edge $(u, v) \in E$ from C, and construct a new graph $G' = (V, E \setminus \{(u, v)\})$ by deleting (u, v). Now, join two copies G_1' and G_2' of G' to obtain a graph G''. Let $f_1 : V(G) \to V(G_1')$ and $f_2 : V(G) \to V(G_2')$ be the functions that map a node $w \in G$ onto its copy in G_1' and G_2', respectively. To simplify notation, we denote the two copies of an arbitrary node $x \in G$ by $x_1 := f_1(x)$ and $x_2 := f_2(x)$, respectively. Insert two edges (u_1, v_2) and (u_2, v_1) into the joined graph G'', and call the resulting graph H (see Fig. 4.2).

The edge expansion of H is $h(G) = \frac{2}{n/2} = 4/n$ because the cut of G_1' in H has size two. Furthermore, G_1' and G_2' are connected because G is connected and only a single edge from a cycle was deleted in G_1' and G_2', respectively. The latter graphs were joined by two edges to obtain H, and therefore H is connected. It remains to prove that $\mathrm{freq}_k(G) = \mathrm{freq}_k(H)$.

Let $r \in G$ be an arbitrary node. We show that $\mathrm{disk}_k(G, r) \cong \mathrm{disk}_k(H, r_1)$. Since H contains a (non-trivial) automorphism, this is equivalent to $\mathrm{disk}_k(G, r) \cong \mathrm{disk}_k(H, r_2)$. We define a mapping $f : V(\mathrm{disk}_k(G, r)) \to V(\mathrm{disk}_k(H, r_1))$ as follows. Let $x \in \mathrm{disk}_k(G, r)$. If $x_1 \in \mathrm{disk}_k(H, r_1)$, then define $f(x) := x_1$. Otherwise, set $f(x) := x_2$. Clearly, f is injective. To see that f is surjective, observe that

$$x \in \mathrm{disk}_k(G, r) \Leftrightarrow x_1 \in \mathrm{disk}_k(H, r_1) \vee x_2 \in \mathrm{disk}_k(H, r_1)$$

because for every edge $(y, z) \in G$ there exist

$$\begin{array}{rll} \text{either} & (y_1, z_2), (y_2, z_1) \in H & \text{if } (y, z) = (u, v) \\ \text{or} & (y_1, z_1), (y_2, z_2) \in H & \text{otherwise.} \end{array}$$

In other words, all paths from the root of $\mathrm{disk}_k(G, r)$ to one of its nodes x can be mapped to a path from the root of $\mathrm{disk}_k(H, r_1)$ to x_1 or x_2 of the same length and vice versa. However, x_1 and x_2 are not contained in $\mathrm{disk}_k(H, r_1)$ simultaneously. Actually, the following stronger version of this statement, which we will prove later, holds.

Claim 4.13. *Let* $(x, y) \in \{(z, z) \mid z \in V(G)\} \cup E(G) \setminus \{(u, v)\}$. *It holds that* $x_1 \notin \mathrm{disk}_k(H, r_1)$ *or* $y_2 \notin \mathrm{disk}_k(H, r_1)$.

We show that $(x, y) \in \mathrm{disk}_k(G, r) \Leftrightarrow (f(x), f(y)) \in \mathrm{disk}_k(H, r_1)$ next. By the above reasoning, we have already shown that

$$(x, y) \in \mathrm{disk}_k(G, r) \Rightarrow (f(x), f(y)) \in \mathrm{disk}_k(H, r_1) \quad \forall (x, y) : \min(\mathrm{d}(r, x), \mathrm{d}(r, y)) < k.$$

We conclude that this is also true for all *solely induced* edges, i.e., $(x, y) \in \mathrm{disk}_k(G, r)$ where $\mathrm{d}(r, x) = \mathrm{d}(r, y) = k$ because there is no such edge in any k-disk of G: it would induce a cycle of length at most $2k + 1$.

Conversely, let $(f(x), f(y)) \in \mathrm{disk}_k(H, r_1)$ be an arbitrary edge. In $\mathrm{disk}_k(H, w_1)$, there exist two paths $P_{f(x)}$ and $P_{f(y)}$ from r_1 to $f(x)$ and $f(y)$, respectively. These paths can be mapped to two paths P_x and P_y of the same length in $\mathrm{disk}_k(G, r)$ by the construction of f. Therefore, we have that $(x, y) \in \mathrm{disk}_k(G, r)$.

It follows that, for every k-disk $\mathrm{disk}_k(G, r))$, we have $\mathrm{disk}_k(G, r)) \cong \mathrm{disk}_k(H, r_1)) \cong \mathrm{disk}_k(H, r_2))$. Therefore, $\mathrm{freq}_k(H) = 2 \cdot \mathrm{dist}_k(G) / (2n) = \mathrm{freq}_k(G)$. □

Proof of Claim 4.13. Assume that $x_1 \in \mathrm{disk}_k(G, r)$. The distance of u_2 and v_2 in H is at least $2k + 1$ by Lemma 4.10 (otherwise, there would exist a cycle of length at most $2k + 1$ in G). Therefore, there exists a path P_{x_1} from r_1 to x_1 in $\mathrm{disk}_k(H, r_1)$ that contains neither (u_1, v_2) nor (u_2, v_1) (see Fig. 4.3 on the next page). This implies that there exists a path P_x from r to x of length at most k in $\mathrm{disk}_k(G, r)$ that does not contain (u, v). Now, assume that $y_2 \in \mathrm{disk}_k(H, r_1)$, too. This implies that there exists another path P_{y_2} from r_1 to y_2 of length at most k in $\mathrm{disk}_k(H, r_1)$ that contains (u_1, v_2) or (u_2, v_1). Again, this implies that there exists a path P_y in G in $\mathrm{disk}_k(G, r)$ that contains (u, v). The two paths P_x and P_y induce a cycle of length at most $2k + 1$ because their start points are equal and their end points are equal or adjacent. By Lemma 4.10, we have that the girth of G is at least $2k+2$, which yields a contradiction. □

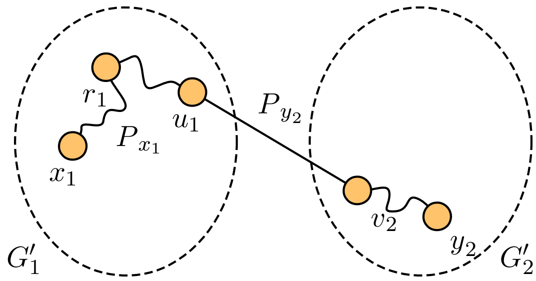


Figure 4.3.: The path P_x, which connects r_1 and x_1, and the path P_y, which connects r_1 and y_2 (in this example, the edge (u_1, v_2) is contained in P_y).

Combining Lemmas 4.8 and 4.12, we conclude that there exists a sequence of graphs that contains, for sufficiently large and even $n \in \mathbb{N}$, a graph H_n such that $|H_n| \geq n$, $\text{freq}_k(H_n, \Delta_{d,k}) = 1$ and $h(H_n) \leq 4/|H_n|$.

Lemma 4.9 (repeated). *For every $d \geq 3$ and $k \geq 1$ there exists a sequence of graphs $(H_i)_{i\in\mathbb{N}}$ such that $|H_i| = 2\nu + 4i$ (where $\nu := \nu(k)$ is defined as in Lemma 4.8), $\text{freq}_k(H_i, \Delta_{d,k}) = 1$ and $h(H_i) = 4/|H_i|$ for every $i \in \mathbb{N}$.*

Proof. Let $i \in \mathbb{N}$ and H'_i be a graph of size $\nu + 2i$ obtained from applying Lemma 4.8, i.e., $\text{freq}_k(H'_i, \Delta_{d,k}) = 1$ and $h(H'_i) > 0$. It is connected because it has edge expansion greater than zero, and it contains at least one cycle because it is finite and d-regular. Therefore, the graph H'_i meets the requirements of Lemma 4.12, and we can use it to obtain a graph H_i of size $2 \cdot (\nu + 2i) = 2\nu + 4i$ with the desired properties. □

4.2. Feasibility of k-Disks

The result of the preceding section poses an interesting question that is not necessarily related to expander graphs. Consider a finite tree T: a k-disk centered at an inner node and a k-disk centered at a leaf are not isomorphic. However, as shown by Lemma 4.8, there exist finite graphs such that all k-disks in these graphs are trees and pairwise isomorphic. This leads to the more general question which k-disks can be the only k-disk that is present in the limit of a graph sequence. In other words, given a k-disk Δ, does there exist a graph sequence $(G_i)_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty} \text{freq}_k(G_i, \Delta) = 1$? For example, the most frequent k-disk in large grids satisfies this (see Section 1.1). However, we do not restrict ourselves to planar or hyperfinite graphs but consider k-disks of all degree-bounded graphs. A related problem is to describe the set of (normalized) k-disk vectors of all degree-bounded graphs, i.e., the maximal set of vectors where every element is the k-disk vector of at least one degree-bounded graph. We denote the set of normalized k-disk vectors of all d-bounded graphs by $\mathcal{D}_{d,k}$, i.e., $\mathcal{D}_{d,k} := \{\mathrm{freq}_k(G) \mid G \text{ is a } d\text{-bounded graph}\}$. In Sections 4.2.1 and 4.2.2, we discuss some minor results regarding the aforementioned aspects of k-disk vectors and k-disk distributions.

4.2.1. Limits of k-Disk Distributions

Some *special* points on the border of $\mathcal{D}_{d,k}$ are those k-disk vectors where the i-th entry equals one and all other entries equal zero for some i. The following result yields a necessary condition that must hold for all k-disks that can appear as the only non-zero entry in the limit of the normalized k-disk vector of a graph sequence.

Theorem 4.14. *Let $d, k \geq 0$ and Δ be a d-bounded k-disk. If there exists a d-bounded graph sequence $(G_i)_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty} \mathrm{freq}_k(G_i, \Delta) = 1$, then there exists a rooted d-bounded graph sequence $(H_i, r_i)_{i\in\mathbb{N}}$ such that the k-disks of all nodes $v \in H_i$ with distance to r_i at most i are isomorphic to Δ for every $i \in \mathbb{N}$.*

Proof. Assume that there exists some $j \in \mathbb{N}$ such that there exists no rooted graph (H_j, r_j) where all nodes with distance to r_j at most j are isomorphic to Δ. We prove that there exists $c > 0$ such that $\mathrm{freq}_k(G, \Delta) \leq 1 - c$ for every d-bounded graph G. Therefore, there exists no d-bounded graph sequence $(G_i)_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty} \mathrm{freq}_k(G_i, \Delta) = 1$.

Let $G = (V, E)$ be a d-bounded graph. By the preceding assumption, there exists a function $f : V \to V$ that maps a node u to a node v such that $\mathrm{disk}_k(G, v) \not\cong \Delta$ and the distance between u and v is at most j. For every $v \in V$, the number of nodes that map to v is at most $3d^j/2$ by Observation 3.8 on page 35. Therefore, the image of f is of size at least $2n/(3d^j)$. It follows that $\mathrm{freq}_k(G, \Delta) \leq 1 - 2/(3d^j) \leq 1 - c$ where $c = 2/(3d^j) > 0$. □

Consider an $n \times n$ grid graph $G_{n\times n}$. Adding a *border around the border* of $G_{n\times n}$ yields an $(n + 2) \times (n + 2)$ grid graph. Since this newly added border is small in relation to the whole graph, the fraction of k-disks whose roots have distance at least k to the border – and therefore the fraction of pairwise isomorphic k-disks in the graph – is increased. The border expansion of a rooted graph (G, r) on level i is the fraction of nodes with distance i to r among all nodes with distance at most i to r.

Definition 4.15 (Border expansion). Let (G, r) be a rooted graph. Denote the size of the i-th level of G by $n(i) := |\{u \in G \mid \mathrm{d}(r, u) = i\}|$. The border expansion of

G on level ℓ is defined as

$$b_G(\ell) = \frac{n(\ell)}{\sum_{i=0}^{\ell-1} n(i)} \,. \qquad \blacksquare$$

This leads to the following sufficient condition, which is based on a small border expansion of a graph sequence, e.g., as it is considered in Theorem 4.14.

Theorem 4.16. *Let $d, k \geq 0$ and Δ be a d-bounded k-disk. If there exists a rooted d-bounded graph sequence $(H_i, r_i)_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty} b_{H_i}(i+1) = 0$ and, for every $i \in \mathbb{N}$, all nodes $u \in H_i$ with $\mathrm{d}(r_i, u) \leq i$ are isomorphic to Δ, then there exists a d-bounded graph sequence $(G_i)_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty} \mathrm{freq}_k(G_i, \Delta) = 1$.*

Proof. For every graph H_i, we construct a graph G_i from H_i by removing all nodes that have distance to r_i at least $i + k + 1$. Let $n_i(j) := |\{u \in G_i \mid \mathrm{d}(r_i, u) = j\}|$ denote the size of the j-th level of G_i. By Observation 3.8 on page 35, at most $3d^{(k-1)}/2 \cdot n_i(i+1)$ nodes have distance at least $i+1$ to r_i. The k-disks of all nodes that have distance at most i to r_i are isomorphic to Δ. Therefore, the k-disks that are potentially not isomorphic to Δ, $\sum_{j=i+1}^{i+k} n_i(j)$, are outnumbered by k-disks that are surely isomorphic to Δ, $\sum_{j=0}^{i} n_i(j)$, as i goes to infinity.

$$\lim_{i\to\infty} \frac{\sum_{j=i+1}^{i+k} n_i(j)}{\sum_{j=0}^{i} n_i(j)} \leq \lim_{i\to\infty} \frac{3d^{(k-1)}}{2} b_{G_i}(i+1) = \lim_{i\to\infty} \frac{3d^{(k-1)}}{2} b_{H_i}(i+1) = 0 \,. \qquad \square$$

Even if a k-disk Δ satisfies Theorem 4.16, there is not necessarily an intuitive way to grow it like there is for k-disks that are subgraphs of grid graphs. For example, recall the k-disk $\Delta_{d,k}$:

Definition 4.6 (repeated). The rooted tree where the distance from the root to each leaf is k and each inner node is d-regular is denoted by $\Delta_{d,k}$. $\blacksquare$

It is tempting to consider the graph sequence $(H_i)_{i\in\mathbb{N}}$ where $H_i := \Delta_{d,i+k}$ because it meets the requirements of Theorem 4.14. However, if $d \geq 3$, the number of leaves of H_i is at least $|H_i|/2$ for every $i \in \mathbb{N}$, which does not sustain Theorem 4.16. This implies that $\lim_{i\to\infty} \mathrm{freq}_k(G_i, \Delta_{d,k}) < 1$. On the other hand, Lemma 4.8 states that there exists a finite graph G such that even $\mathrm{freq}_k(G, \Delta_{d,k}) = 1$ holds.

4.2.2. The Set of k-Disk Vectors

It is not too obvious how to describe $\mathcal{D}_{d,k}$ explicitly. The following result can, however, be used to compute an explicit approximation of $\mathcal{D}_{d,k}$ as a polytope. It states that the set of all k-disk vectors is ϵ-dense in its convex hull for every $\epsilon > 0$.

Lemma 4.17. *Let $\epsilon > 0, \delta \in [0,1], d \geq 0, k \geq 0$ and G_1, G_2 be two d-bounded graphs. Then, there exists a d-bounded graph H such that*

$$\Big|\Big|\mathrm{freq}_k(H) - \big[\delta \cdot \mathrm{freq}_k(G_1) + (1-\delta) \cdot \mathrm{freq}_k(G_2)\,\big]\Big|\Big|_1 \leq \epsilon\,.$$

Proof. This proof is a simplified version of the proof of Lemma 3.7 on page 33, which employs the same idea. Let $\mathcal{F}$ be the family of all graphs and set $N := N_{\mathcal{F}}(d,k)$. Choose $c \in \mathbb{N}$ such that $c \geq 2N/\epsilon$. Join $\lfloor \delta c \rfloor \cdot |G_2|$ copies of G_1 with $\lceil (1-\delta)c \rceil \cdot |G_1|$ copies of G_2 and denote the resulting graph by H. We have that

$$\begin{aligned}
\mathrm{freq}_k(H) &= \frac{1}{c \cdot |G_1| \cdot |G_2|} \cdot \Big(\lfloor \delta c \rfloor \cdot |G_2| \cdot \mathrm{dist}_k(G_1) + \lceil (1-\delta)c \rceil \cdot |G_1| \cdot \mathrm{dist}_k(G_2)\Big) \\
&= \frac{\lfloor \delta c \rfloor}{c} \cdot \mathrm{freq}_k(G_1) + \frac{\lceil (1-\delta)c \rceil}{c} \cdot \mathrm{freq}_k(G_2)\ .
\end{aligned}$$

Therefore, it holds that

$$\begin{aligned}
&\Big|\Big|\mathrm{freq}_k(H) - \big[\delta \cdot \mathrm{freq}_k(G_1) + (1-\delta) \cdot \mathrm{freq}_k(G_2)\,\big]\Big|\Big|_1 \\
\leq\ &\left|\left|\left(\frac{\lfloor \delta c \rfloor}{c} - \delta\right) \cdot \mathrm{freq}_k(G_1) + \left(\frac{\lceil (1-\delta)c \rceil}{c} - (1-\delta)\right) \cdot \mathrm{freq}_k(G_2)\right|\right|_1 \\
\leq\ &\left|\left|\frac{1}{c} \cdot (\mathrm{freq}_k(G_1) + \mathrm{freq}_k(G_2))\right|\right|_1 \leq 2N/c \leq \epsilon\,.
\end{aligned}$$

□

To obtain an explicit approximation of $\mathcal{D}_{d,k}$, one can compute the k-disk vectors of all d-bounded graphs of size at most $M_{\mathcal{F}}(\epsilon, d, k)$, where $\mathcal{F}$ is the family of all graphs, and construct its convex hull, which is a polytope. In Chapter 5, we discuss another possible approach that yields an approximation of $\mathcal{D}_{d,k}$.

4.3. Experimental Analysis

While previous sections dealt with theoretical results about k-disks and k-disk distributions, this section is about the results of experiments conducted to examine k-disks in real graph instances. As mentioned in Chapter 1, k-disks often serve as a means to an end, e.g., in property testing. Besides, k-disks constitute a single node's view on its local surroundings by itself. For example, it might be helpful to sample k-disks and analyze them to gain knowledge of friendship relations in a social network.

Without any further steps taken, this approach is particularly promising if k is small such that there is a good chance that some k-disks appear often in the graph, say, for more than 5% of its nodes. Therefore, the experiments in this section are carried out with the following questions in mind:

1. *Does the graph have frequent k-disks?* For example, one might conjecture that about 75% of the 1-disks in a road network graph represent 4-way intersections (nodes with degree four). On the other hand, it is not clear whether there are also 2- or 3-disks with a frequency of more than, say, 1%.

2. *What informations can be obtained from the k-disk vector?* The answer to this question is very specific to the actual graph. However, it should give an idea of the type of information one may obtain by examining local neighborhoods of a graph.

4.3.1. Setup

For each graph, the following experiment is conducted for various k and sample size σ: A total of σ nodes is sampled from the input graph and their k-disks are determined. For each isomorphism type, the number of occurrences is recorded.

We examine all k-disks Δ that have a frequency of at least 5%, i.e., at least 5% of the sampled k-disks are isomorphic to Δ. Ideally, one would calculate the frequency of each k-disk from the whole input graph to answer the questions above. However, examining large graphs and counting their k-disk isomorphism types places high demands on memory and computing capacities. Most graph libraries require that a graph is loaded into the main memory before even simple operations (e.g., querying the existence of an edge) can be performed. Additionally, all k-disk isomorphism types that were encountered so far have to be stored.

Checking whether a sampled k-disk is isomorphic to a previously sampled one can be an expensive operation. Although running an algorithm to check if two k-disks are isomorphic can often be avoided (e.g., by comparing the number of nodes that have distance ℓ to the root for all $\ell \leq k$), it is unavoidable sometimes. Since the computational complexity of graph isomorphism has not been determined in general yet (it is not known whether it belongs to P or is NP-complete), there are no efficient algorithms known. Therefore, it is difficult to examine and compare all k-disks in a graph. The results in this section have been obtained by using the Boost Graph Library[†], which is a well-established graph library that provides an implementation of a graph isomorphism algorithm.

We present results for six graphs provided by SNAP [34]:

[†] `http://www.boost.org/`

Graph	Number of nodes	Number of edges
com-Amazon	$3.3 \cdot 10^5$	$9.3 \cdot 10^5$
com-DBLP	$3.2 \cdot 10^5$	$1.0 \cdot 10^6$
com-YouTube	$1.1 \cdot 10^6$	$3.0 \cdot 10^6$
roadNet-CA	$2.0 \cdot 10^6$	$2.8 \cdot 10^6$
roadNet-PA	$1.1 \cdot 10^6$	$1.5 \cdot 10^6$
roadNet-TX	$1.4 \cdot 10^6$	$1.9 \cdot 10^6$

Table 4.1.: Graphs from [34] used in the experiments

com-Amazon is a co-purchasing network obtained from Amazon‡. Each node represents a product. Two nodes are connected iff they are frequently purchased together.

com-DBLP is a co-authorship network graph obtained from the computer science bibliography DBLP§. Each node represents an author of a publication. Two nodes are connected iff the corresponding authors have published together.

com-YouTube is a friendship network obtained from YouTube¶. Each node represents a member of YouTube. Two nodes are connected iff the corresponding members have declared their friendship.

road-CA, road-PA and road-TX are the road networks of California, Pennsylvania and Texas, respectively. Each node represents an intersection or an endpoint of a road. Two nodes are connect iff there exists a road connecting these intersections or endpoints, respectively.

All graphs are connected. See Table 4.1 for the number of nodes and edges, respectively. The maximum node degrees of the datasets com-Amazon, com-DBLP and com-YouTube are not bounded by a small constant, e.g., less than 100. On the other hand, the degrees of the road networks road-CA, road-PA and road-TX are bounded because intersections where more than, say, six roads meet are unfavorable. Moreover, road networks are planar except for a few tunnels and bridges. Therefore, these graphs can be safely considered as being hyperfinite and almost planar.

We give results for $k \in \{1, 2, 3\}$ for all graphs except com-YouTube. Experiments for com-YouTube were only performed for $k \in \{1, 2\}$ due to limited memory (8GB main memory). The sample size is $\sigma = 1\,000$. Therefore, the probability that the

‡`http://www.amazon.com/`
§`http://dblp.uni-trier.de/`
¶`http://www.youtube.com/`

(measured) frequency of a k-disk that is centered at 5% (50%) of all nodes is between 3% and 7% (45% and 55%) is more than 99.5% by the p-value of the corresponding Bernoulli experiment.

4.3.2. Results

In addition to the condensed information depicted by the figures in this section, more details can be found in Appendix A.

Frequent k-Disks

For $k = 1$, the sample sets of all graphs contain k-disks whose frequency is at least 5% (see Fig. 4.4 on page 64 and Table 4.2). However, the distribution of (measured) probability mass among the most frequent k-disks differs. About half of the k-disks sampled from com-YouTube and the road networks are isomorphic. On the other hand, half of the probability mass is scattered across the 14 and 4 most frequent k-disks of com-Amazon and com-DBLP, respectively.

For $k = 2$, only com-YouTube and the road networks have frequently sampled k-disks (see Fig. 4.5a and Table 4.2). The top three k-disks bind about 10% and 20% of the total probability mass, respectively. However, the remaining probability mass spreads widely. The significance of the most frequent k-disks is even smaller for $k = 3$, where 10% of the probability mass is distributed to the approximately four most frequent k-disks of the road networks (see Fig. 4.5b). All 3-disks that were sampled from com-Amazon and com-DBLP have a frequency of less than 1%.

Information Retrieved From Frequent k-Disks

The following information can be read directly from the most frequent 1-disks without any further efforts (see Table 4.2). About 7% of the products from the com-Amazon dataset are often co-purchased with exactly one other product, another 7% are co-purchased with two other products, which however are not frequently co-purchased. Approximately 16% of the authors from com-DBLP have published papers with one other author only. 14% (13%, 7%, 4%) have only published papers with the same 2 (3, 4, 5) co-authors and possibly some papers with a subset of them. Most users (54%) from com-YouTube have listed only one friend and about 11% have listed (exactly) two friends. Almost half of the junction points (42%) in road-CA are T-junctions, i.e., 3-way intersections. 15% are 4-way intersections and another 15% are end points. Among the 2-disks, 12% of the junction points are end points of dead ends that are connected to another street by a T-junction. There are no 3-disks that account for more than 5% of the sampled nodes. The results for road-PA and road-TX are similar.

4.3.3. Evaluation

All tested graphs contain frequent 1-disks. However, the number of such k-disks as well as their frequency decreases for $k = 2$ and $k = 3$ by comparison to the case $k = 1$. Even road networks, which are composed of a small number of local structures, do not feature any 3-disks that were sampled with a frequency of more than 5%. The present results raise the question whether the probability mass that can be assigned to a single k-disk is bounded for some (common) families of graphs. This question is not answered by the experiments and may be subject to further studies.

In summary, some information about the very local structure of the tested graphs can be obtained from the sampled 1-disks and their frequency *for free*. However, even for 2-disks and the planar road networks this informations is limited. This limitation is likely due to the fragility of isomorphism: changing a single edge can already break isomorphism between two graphs. One can possibly overcome this problem by defining a graph parameter function, i.e., a (typically non-injective) function that maps each k-disk to a feature vector. For example, one can map a k-disk to the number of nodes with distance three to the root and apply this function to the k-disks sampled from com-YouTube to examine the number of *friends of third degree*. However, defining such a function requires some idea of what shall be analyzed, while interpreting sampled k-disks does not.

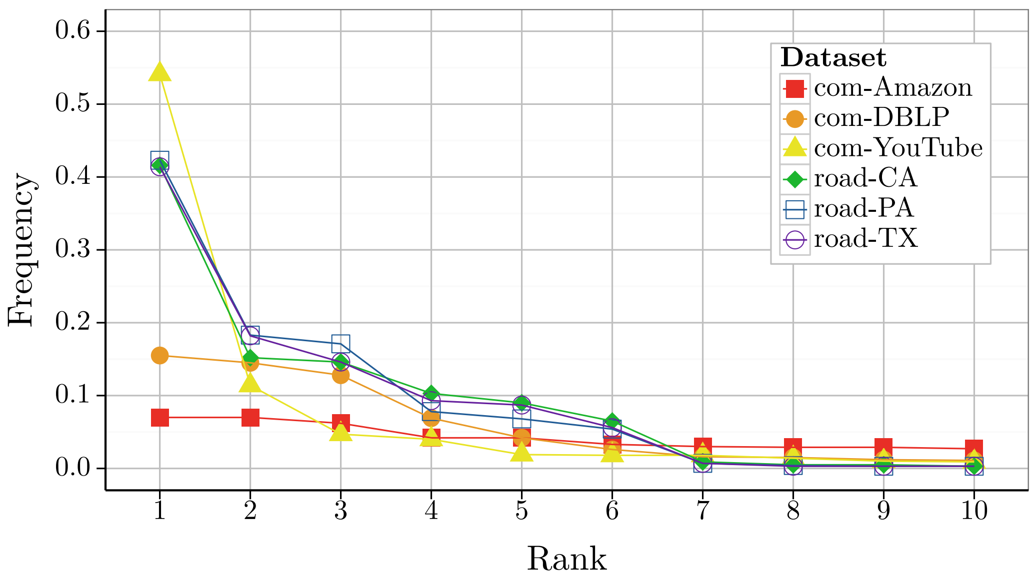
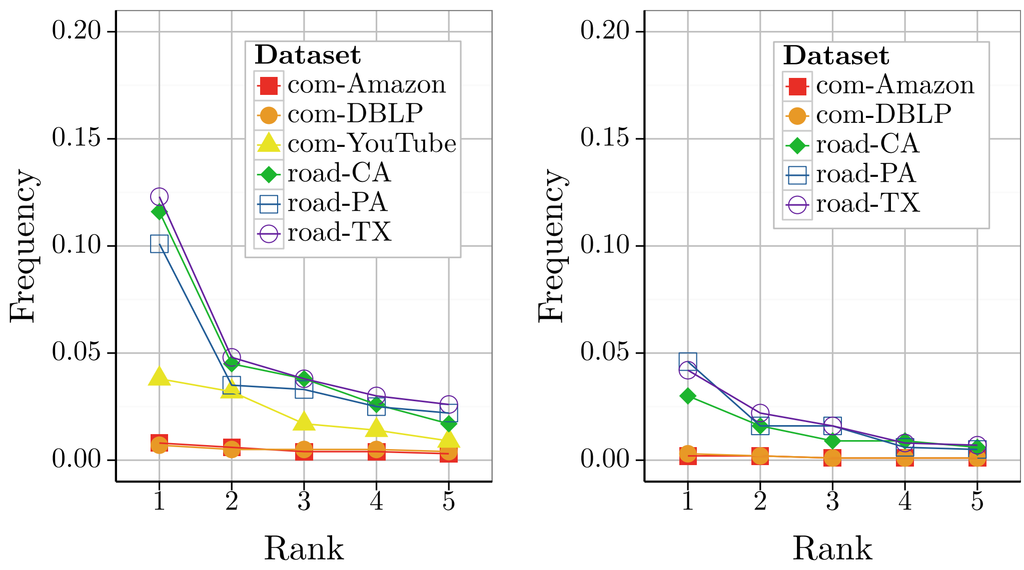


Figure 4.4.: Ranks of the ten most frequent 1-disks that were sampled from each dataset are plotted against their actual frequencies. Lines connecting the points are provided for visual guidance.



(a) 2-disks

(b) 3-disks

Figure 4.5.: Ranks of the five most frequent (a) 2-disks and (b) 3-disks that were sampled from each dataset are plotted against their actual frequencies. Lines connecting the points are provided for visual guidance.

	1-disks			2-disks			
	freq.	edg.	lvl.1	freq.	edg.	lvl.1	lvl.2
com-Amazon	0.07	1	1	0.01	3	1	2
	0.07	2	2	0.01	2	1	1
	0.06	4	3	< 0.01	4	1	2
	0.04	3	3	< 0.01	4	1	3
	0.04	3	2	< 0.01	6	1	3
com-DBLP	0.16	1	1	0.01	2	1	1
	0.14	3	2	< 0.01	7	2	3
	0.13	6	3	< 0.01	9	2	3
	0.07	10	4	< 0.01	4	1	2
	0.04	15	5	< 0.01	7	3	1
com-YouTube	0.54	1	1	0.04	2	1	1
	0.11	2	2	0.03	3	1	2
	0.05	3	3	0.02	4	1	3
	0.04	3	2	0.01	5	1	4
	0.02	5	5	0.01	12	1	11
road-CA	0.42	3	3	0.12	3	1	2
	0.15	4	4	0.04	4	1	3
	0.15	1	1	0.04	7	3	4
	0.10	2	2	0.03	9	3	6
	0.09	4	3	0.02	6	2	4
road-PA	0.42	3	3	0.10	3	1	2
	0.18	1	1	0.04	7	3	4
	0.17	4	4	0.03	4	1	3
	0.08	2	2	0.02	9	3	6
	0.07	4	3	0.02	8	3	5
road-TX	0.41	3	3	0.12	3	1	2
	0.18	1	1	0.05	7	3	4
	0.15	4	4	0.04	9	3	6
	0.09	4	3	0.03	4	1	3
	0.09	2	2	0.03	8	3	5

Table 4.2.: The five most frequent 1- and 2-disks in a sample of size 1 000 drawn from each dataset uniformly at random. For each k-disk, the frequency (freq.), the number of edges (edg.) and the number of nodes that have distance x to the root (lvl.<x>) are stated. The k-disk are sorted according to their frequency (ties broken arbitrarily). If „< 0.01“ is indicated in the frequency column, the value would have been rounded to zero. See Appendix A for more details and the case $k = 3$.

5. Discussion

In this thesis, we have discussed various results related to graphs and their k-disks. The main result is related to a conjecture by Lovász, which has been proved by Alon: for every d-bounded graph G, $k \in \mathbb{N}$ and $\epsilon > 0$, there exists a small graph H such that $|H|$ is constant in terms of $|G|$ and $||\text{freq}_k(G) - \text{freq}_k(H)\,||_1 < \epsilon$. However, the proof does not imply an effective bound on the size of H.

We have given explicit upper bounds of order $\mathcal{O}(d^{3k+2}\, 2^{7.5d^k}/\epsilon^4)$ on the size of H if G is planar in Chapter 3. Moreover, we have generalized this bound to $\mathcal{O}(N_{\mathcal{F}}(d,k) \cdot \varphi(\epsilon \,/\, (9d^k)) \,/\, \epsilon^2)$ for any family $\mathcal{F}$ of $\varphi(\cdot)$-hyperfinite graphs. Since not all graphs are hyperfinite in practice (e.g., it is preferable that communication networks are not hyperfinite), it is a major open problem to find such bounds for other types of graphs as well. A natural enhancement would be to prove bounds for expander graphs. Expander graphs are somewhat dual to hyperfinite graphs: they are, in contrast, globally connected and cannot be decomposed into relatively independent subgraphs. Therefore, it seems very unlikely that the idea of the proof for planar graphs, i.e., splitting the graph into small components by removing only a small fraction of its edges, can be extended to expander graphs. Nevertheless, it might be possible to combine bounds on hyperfinite graphs and expanders graphs to obtain bounds for (more) general graphs by distinguishing between hyperfinite and expander parts.

As mentioned earlier, k-disk vectors describe the distribution of local neighborhoods of a graph. It has been proved by Newman and Sohler [44] that two planar graphs with similar k-disk vectors can be transformed into each other by changing only a constant fraction of their edges. Focusing on a single graph, we have given a lower bound on k when a graph of size n is encoded by a k-disk vector and ϵn edge operations for a restricted choice of ϵ. The restriction is imposed by the encoding argument we have used. A valuable improvement would be to eliminate this restriction, i.e., prove the result for $\epsilon \in (0, 1]$ for every d-bounded planar graph. Again, it would also be interesting to extend this result to other families of graphs such as hyperfinite – or even more general – graphs.

In Chapter 4, we have analyzed the relation between graphs and their k-disks. First, we have directed our attention to the k-disks of some graphs that contain no short cycles and these graph's expansion. We have shown that there exist graphs with good and weak expansion properties that share the same k-disk vector. In other words, the k-disk vector of a degree-bounded graph G contains only very little information about the expansion of G in general. However, there remain many things regarding expander graphs and their k-disks one might be interested in, e.g., which k-disks can only appear rarely (e.g., subgraphs of grids) or if most k-disks of expander graphs must fulfill certain conditions.

In a more general context, we have proved some minor results regarding the set of d-bounded k-disk vectors. Finding a way to compute an approximation of (a part of) this set would help to construct an algorithm that, given a k-disk vector $\vec{\boldsymbol{u}}$ and $\epsilon > 0$, decides if there exists a graph G such that $||\text{freq}_k(G) - \vec{\boldsymbol{u}}||_1 < \epsilon$ (and possibly constructs G if it exists). For example, one might come up with an algorithm that, given a vector $\vec{\boldsymbol{s}} \in \mathbb{R}^N$, constructs a graph G that minimizes $\langle \vec{\boldsymbol{s}}, \text{freq}_k(G) \rangle$ and apply scalarization techniques known from multiple objective integer programming: using various vectors $\vec{\boldsymbol{s}}$, such an algorithm can be used to discover nodes on the boundary of a convex set. Since $\mathcal{D}_{d,k}$ is ϵ-dense in its convex hull for every $\epsilon > 0$, these nodes might be used to approximate $\mathcal{D}_{d,k}$ by the intersections of half spaces, i.e., a polytope (see Fig. 5.1 on the following page). An overview of common scalaraization techniques is, e.g., provided by Ehrgott [18].

We have also discussed a necessary and a sufficient condition for k-disks Δ that can appear as the only k-disk in the limit of a graph sequence. The sufficient condition is based on the natural idea of growing a graph such that only the k-disks of new nodes are not isomorphic to Δ. However, to obtain a necessary and sufficient condition from this approach, one needs to deal with possibly large growth rates. Overall, it seems essential to gain a better understanding of how k-disks can be put together to assemble a global structure, e.g., long cycles.

Apart from the aforementioned theoretical results, we have discussed the results of some experiments to analyze the k-disks of some real-world graphs. In particular, we have studied whether the probed networks contain frequent k-disks and, if they do, whether these k-disks contain *useful* (structural) information. Briefly speaking, there exists a set of a couple of 1-disks that are centered at more than five percent of a graph's nodes for all tested graphs. This is not the case for $k = 2$ and every tested graph anymore. For $k \geq 3$, there is no k-disk with a frequency of at least five percent for any of the tested graphs. However, all frequent k-disks that have been found contain some meaningful information about the local structure of the

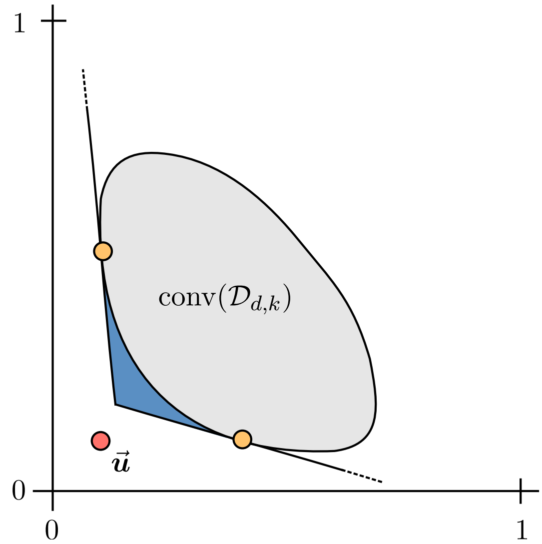


Figure 5.1.: A part of the convex hull of $\mathcal{D}_{d,k}$ is approximated by a polytope to test whether $\vec{\boldsymbol{u}}$ is close to a feasible k-disk vector. The testing algorithm answers yes iff $\vec{\boldsymbol{u}}$ lies inside of the polytope. The answer is wrong only if $\vec{\boldsymbol{u}}$ lies in the blue area between the border of conv($\mathcal{D}_{d,k}$) and the border of the polytope.

data that is represented by its corresponding graph. For example, the majority of the authors from the com-DBLP dataset has published only within one group of coauthors, i.e., the same people, which have published all together for at least one time. One can extend this approach by, e.g., using graph parameter functions to extract informations from k-disk for large(r) k. Therefore, k-disks might serve a purpose similar to other models of local structure, e.g., motifs [43].

A. Results of the Experiments

The following tables provide details about the results of the experiments described in Section 4.3. For each k-disk, the proportion of the whole sample set (freq.), its number of edges (edg.) and the number of nodes with distance x to the root (lvl.<x>) are indicated. The k-disk are sorted according to their frequency (ties broken arbitrarily). If „< 0.01“ is indicated in the frequency column, the value would have been rounded to zero.

com-Amazon											
1-disks			2-disks				3-disks				
freq.	edg.	lvl.1	freq.	edg.	lvl.1	lvl.2	freq.	edg.	lvl.1	lvl.2	lvl.3
0.07	1	1	0.01	3	1	2	< 0.01	389	2	51	141
0.07	2	2	0.01	2	1	1	< 0.01	37	5	3	9
0.06	4	3	< 0.01	4	1	2	< 0.01	1 197	17	104	500
0.04	3	3	< 0.01	4	1	3	< 0.01	155	6	6	51
0.04	3	2	< 0.01	6	1	3	< 0.01	73	4	9	19
0.03	5	3	< 0.01	12	2	5	< 0.01	88	11	9	13
0.03	6	4	< 0.01	5	1	4	< 0.01	77	10	11	3
0.03	5	4	< 0.01	5	2	3	< 0.01	243	5	23	78
0.03	7	4	< 0.01	5	3	1	< 0.01	636	3	51	256
0.03	6	3	< 0.01	4	2	1	< 0.01	670	5	16	313
0.02	10	4	< 0.01	7	3	4	< 0.01	113	3	19	35
0.02	8	4	< 0.01	7	1	5	< 0.01	375	1	47	72
0.02	9	4	< 0.01	71	5	25	< 0.01	43	2	7	7
0.02	14	5	< 0.01	5	1	3	< 0.01	838	3	103	185
0.02	4	4	< 0.01	24	6	2	< 0.01	143	4	11	53
0.01	6	5	< 0.01	51	8	20	< 0.01	35	1	7	6
0.01	13	5	< 0.01	35	5	11	< 0.01	817	5	53	313
0.01	15	5	< 0.01	78	15	30	< 0.01	62	2	11	21
0.01	9	5	< 0.01	37	3	13	< 0.01	103	5	16	31
0.01	6	4	< 0.01	154	7	76	< 0.01	129	4	22	39

Table A.1.: The twenty most frequent k-disks in a sample of size 1 000 drawn from com-amazon uniformly at random.

com-DBLP											
1-disks			2-disks				3-disks				
freq.	edg.	lvl.1	freq.	edg.	lvl.1	lvl.2	freq.	edg.	lvl.1	lvl.2	lvl.3
0.16	1	1	0.01	2	1	1	< 0.01	860	1	35	332
0.14	3	2	< 0.01	7	2	3	< 0.01	15	5	1	1
0.13	6	3	< 0.01	9	2	3	< 0.01	6 370	17	139	1 805
0.07	10	4	< 0.01	4	1	2	< 0.01	957	1	30	335
0.04	15	5	< 0.01	7	3	1	< 0.01	35	2	11	12
0.03	21	6	< 0.01	7	1	3	< 0.01	2 503	2	36	474
0.02	28	7	< 0.01	6	1	4	< 0.01	4 615	9	100	1 245
0.02	4	3	< 0.01	11	1	4	< 0.01	198	2	35	71
0.01	5	3	< 0.01	15	1	7	< 0.01	777	6	26	169
0.01	2	2	< 0.01	8	1	4	< 0.01	799	2	24	253

Table A.2.: The ten most frequent k-disks in a sample of size 1 000 drawn from com-DBLP uniformly at random.

com-YouTube						
1-disks			2-disks			
freq.	edg.	lvl.1	freq.	edg.	lvl.1	lvl.2
0.54	1	1	0.04	2	1	1
0.11	2	2	0.03	3	1	2
0.05	3	3	0.02	4	1	3
0.04	3	2	0.01	5	1	4
0.02	5	5	0.01	12	1	11
0.02	4	3	0.01	6	1	5
0.02	4	4	0.01	8	1	7
0.01	5	4	0.01	10	1	9
0.01	7	6	0.01	21 144	1	9 761
0.01	6	4	< 0.01	8	1	6

Table A.3.: The ten most frequent k-disks in a sample of size 1 000 drawn from com-YouTube uniformly at random.

road-CA										
1-disks			2-disks				3-disks			
freq.	edg.	lvl.1	freq.	edg.	lvl.1	lvl.2	freq.	edg.	lvl.1	lvl.2	lvl.3
0.42	3	3	0.12	3	1	2	0.03	7	1	2	4
0.15	4	4	0.04	4	1	3	0.02	5	1	2	2
0.15	1	1	0.04	7	3	4	0.01	8	1	2	4
0.10	2	2	0.03	9	3	6	0.01	8	1	2	5
0.09	4	3	0.02	6	2	4	0.01	1	1	0	0
0.06	5	4	0.02	10	3	6	< 0.01	9	1	2	5
0.01	6	4	0.01	8	3	4	< 0.01	6	1	2	3
< 0.01	6	4	0.01	5	1	3	< 0.01	36	4	8	12
< 0.01	3	2	0.01	11	3	7	< 0.01	8	1	2	4
< 0.01	5	3	0.01	8	3	5	< 0.01	13	3	4	6

Table A.4.: The ten most frequent k-disks in a sample of size 1 000 drawn from road-ca uniformly at random.

road-CA											
1-disks			2-disks				3-disks				
freq.	edg.	lvl.1	freq.	edg.	lvl.1	lvl.2	freq.	edg.	lvl.1	lvl.2	lvl.3
0.42	3	3	0.12	3	1	2	0.03	7	1	2	4
0.15	4	4	0.04	4	1	3	0.02	5	1	2	2
0.15	1	1	0.04	7	3	4	0.01	8	1	2	4
0.10	2	2	0.03	9	3	6	0.01	8	1	2	5
0.09	4	3	0.02	6	2	4	0.01	1	1	0	0
0.06	5	4	0.02	10	3	6	< 0.01	9	1	2	5
0.01	6	4	0.01	8	3	4	< 0.01	6	1	2	3
< 0.01	6	4	0.01	5	1	3	< 0.01	36	4	8	12
< 0.01	3	2	0.01	11	3	7	< 0.01	8	1	2	4
< 0.01	5	3	0.01	8	3	5	< 0.01	13	3	4	6

Table A.5.: The ten most frequent k-disks in a sample of size 1 000 drawn from road-pa uniformly at random.

road-CA											
1-disks			2-disks				3-disks				
freq.	edg.	lvl.1	freq.	edg.	lvl.1	lvl.2	freq.	edg.	lvl.1	lvl.2	lvl.3
0.42	3	3	0.12	3	1	2	0.03	7	1	2	4
0.15	4	4	0.04	4	1	3	0.02	5	1	2	2
0.15	1	1	0.04	7	3	4	0.01	8	1	2	4
0.10	2	2	0.03	9	3	6	0.01	8	1	2	5
0.09	4	3	0.02	6	2	4	0.01	1	1	0	0
0.06	5	4	0.02	10	3	6	< 0.01	9	1	2	5
0.01	6	4	0.01	8	3	4	< 0.01	6	1	2	3
< 0.01	6	4	0.01	5	1	3	< 0.01	36	4	8	12
< 0.01	3	2	0.01	11	3	7	< 0.01	8	1	2	4
< 0.01	5	3	0.01	8	3	5	< 0.01	13	3	4	6

Table A.6.: The ten most frequent k-disks in a sample of size 1 000 drawn from road-tx uniformly at random.

Bibliography

[1] D. Aldous and R. Lyons. Processes on Unimodular Random Networks. *Electronic Journal of Probability*, 12(54), 2007.

[2] N. Alon. Eigenvalues and Expanders. *Combinatorica*, 6(2), 1986.

[3] N. Alon. On the Edge-Expansion of Graphs. *Combinatorics, Probability and Computing*, 11, 1993.

[4] N. Alon and V. D. Milman. λ_1, Isoperimetric Inequalities for Graphs, and Superconcentrators. *Journal of Combinatorial Theory, Series B*, 38(1), 1985.

[5] N. Alon, P. Seymour, and R. Thomas. A Separator Theorem for Nonplanar Graphs. *Journal of the American Mathematical Society*, 3(4), 1990.

[6] N. Alon, P. Seymour, and R. Thomas. A Separator Theorem for Graphs with an Excluded Minor and Its Applications. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing*, 1990.

[7] N. Alon, E. Fischer, I. Newman, and A. Shapira. A Combinatorial Characterization of the Testable Graph Properties: It's All about Regularity. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, 2006.

[8] I. Benjamini, O. Schramm, and A. Shapira. Every Minor-Closed Property of Sparse Graphs Is Testable. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, 2008.

[9] H. L. Bodlaender, A. Grigoriev, and A. Koster. Treewidth Lower Bounds with Brambles. *Algorithmica*, 51(1), 2008.

[10] N. Bonichon, C. Gavoille, N. Hanusse, D. Poulalhon, and G. Schaeffer. Planar Graphs, via Well-Orderly Maps and Trees. In *Graph-Theoretic Concepts in Computer Science*, 2005.

[11] N. Bonichon, C. Gavoille, and N. Hanusse. An Information-Theoretic Upper Bound on Planar Graphs Using Well-Orderly Maps. In *Towards an Information Theory of Complex Networks*. Springer, 2011.

[12] P. Buser. A Note on the Isoperimetric Constant. In *Annales scientifiques de l'École Normale Supérieure*, volume 15, 1982.

[13] C. Carathéodory. Über den Variabilitätsbereich der Fourier'schen Konstanten von positiven harmonischen Funktionen. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 32(1), 1911.

[14] J. Cheeger. A Lower Bound for the Smallest Eigenvalue of the Laplacian. *Problems in Analysis*, 625, 1970.

[15] F. R. K. Chung and R. L. Graham. On Graphs Which Contain All Small Trees. *Journal of Combinatorial Theory, Series B*, 24(1), 1978.

[16] A. Czumaj and C. Sohler. Testing Expansion in Bounded-Degree Graphs. *Combinatorics, Probability and Computing*, 19(5-6), 2010.

[17] J. Dodziuk. Difference Equations, Isoperimetric Inequality and Transience of Certain Random Walks. *Transactions of the American Mathematical Society*, 284(2), 1984.

[18] M. Ehrgott. A Discussion of Scalarization Techniques for Multiple Objective Integer Programming. *Annals of Operations Research*, 147(1), 2006.

[19] G. Elek. L2-Spectral Invariants and Convergent Sequences of Finite Graphs. *Journal of Functional Analysis*, 254(10), 2008.

[20] G. Elek and G. Lippner. An Analogue of the Szemerédi Regularity Lemma for Bounded Degree Graphs. *arXiv preprint arXiv:0809.2879*, 2008.

[21] J. Friedman. A Proof of Alon's Second Eigenvalue Conjecture. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing*, 2003.

[22] A. Frieze and R. Kannan. Quick Approximation to Matrices and Applications. *Combinatorica*, 19(2), 1999.

[23] O. Goldreich. Basic Facts about Expander Graphs. In O. Goldreich, editor, *Studies in Complexity and Cryptography: Miscellanea on the Interplay between Randomness and Computation*, volume 6650 of *Lecture Notes in Computer Science*. Springer, 2011.

[24] O. Goldreich and D. Ron. Property Testing in Bounded Degree Graphs. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, 1997.

[25] O. Goldreich and D. Ron. A Sublinear Bipartiteness Tester for Bounded Degree Graphs. *Combinatorica*, 19(3), 1999.

[26] O. Goldreich and D. Ron. On Testing Expansion in Bounded-Degree Graphs. 2000.

[27] O. Goldreich, S. Goldwasser, and D. Ron. Property Testing and Its Connection to Learning and Approximation. In *Proceedings of the 37th Annual IEEE Symposium on Foundations of Computer Science*, 1996.

[28] R. Halin. S-Functions for Graphs. *Journal of Geometry*, 8(1-2), 1976.

[29] A. Hassidim, J. A. Kelner, H. N. Nguyen, and K. Onak. Local Graph Partitions for Approximation and Testing. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science*, 2009.

[30] S. Hoory, N. Linial, and A. Wigderson. Expander Graphs and Their Applications. *Bulletin of the American Mathematical Society*, 43(4), 2006.

[31] R. R. Howell. On Asymptotic Notation with Multiple Variables, 2008.

[32] H. Huang, N. Linial, H. Naves, Y. Peled, and B. Sudakov. On the 3-Local Profiles of Graphs. *Journal of Graph Theory*, 76(3), 2014.

[33] P. Indyk, A. McGregor, I. Newman, and K. Onak. Open Problems in Data Streams, Property Testing, and Related Topics, 2011.

[34] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford Large Network Dataset Collection, 2014.

[35] K.-i. Kawarabayashi and B. Reed. A Separator Theorem in Minor-Closed Classes. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, 2010.

[36] C. Kuratowski. Sur le probleme des courbes gauches en topologie. *Fundamenta mathematicae*, 15(1), 1930.

[37] R. Levi and D. Ron. A Quasi-Polynomial Time Partition Oracle for Graphs with an Excluded Minor. In *Proceedings of the 40th International Colloquium on Automata, Languages, and Programming*. Springer, 2013.

[38] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Soc, 2009.

[39] R. J. Lipton and R. E. Tarjan. A Separator Theorem for Planar Graphs, 1977.

[40] R. J. Lipton and R. E. Tarjan. Applications of a Planar Separator Theorem. In *Proceedings of the 18th Annual IEEE Symposium on Foundations of Computer Science*, 1977.

[41] L. Lovász. Very large graphs. *Current Developments in Mathematics*, 2008 (2009), 2009.

[42] L. Lovász. *Large Networks and Graph Limits*, volume 60. American Mathematical Soc, 2012.

[43] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298 (5594), 2002.

[44] I. Newman and C. Sohler. Every Property of Hyperfinite Graphs Is Testable. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, 2011.

[45] D. Poulalhon and G. Schaeffer. Optimal Coding and Sampling of Triangulations. In *Proceedings of the 30th International Colloquium on Automata, Languages, and Programming*. Springer, 2003.

[46] N. Robertson and P. D. Seymour. Graph Minors. III. Planar Tree-Width. *Journal of Combinatorial Theory, Series B*, 36(1), 1984.

[47] E. Steinitz. Bedingt konvergente Reihen und konvexe Systeme. *Journal für die reine und angewandte Mathematik*, 143, 1913.

[48] E. Szemerédi. Regular Partitions of Graphs, 1975.

[49] W. T. Tutte. A Census of Planar Triangulations. *Canadian Journal of Mathematics*, 14(1), 1962.

[50] J. Ugander, L. Backstrom, and J. Kleinberg. Subgraph Frequencies: Mapping the Empirical and Extremal Geography of Large Graph Collections. In *Proceedings of the 22nd International Conference on World Wide Web*, 2013.

[51] P. Ungar. A Theorem on Planar Graphs. *Journal of the London Mathematical Society*, 1(4), 1951.

[52] N. C. Wormald. The Asymptotic Distribution of Short Cycles in Random Regular Graphs. *Journal of Combinatorial Theory, Series B*, 31(2), 1981.

Index