On the isomorphism testing of graphs

Xiaorui Sun

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

On the isomorphism testing of graphs

Xiaorui Sun

Graph Isomorphism is one of the very few classical problems in NP of unsettled complexity status. The families of highly regular structures, for example Steiner 2-designs, strongly regular graphs and primitive coherent configurations, have been perceived as difficult cases for graph isomorphism. These highly regular structures arise naturally as obstacles for both the classical group theory and combinatorial approaches for the graph isomorphism problem.

In this thesis we investigate the isomorphism problem of highly regular structures. We present new results to understand the combinatorial structure of highly regular structures, and propose some new algorithms to compute the canonical forms (and thus isomorphism testing) of highly regular structures based on the structural theorems.

We also give an algorithm solving the isomorphism problem of two unknown graphs in the property testing setting. Our new algorithm has sample complexity matching the information theoretical lower bound up to some multiplicative subpolynomial factor.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am truly grateful to my advisor Xi Chen, for spending lots of time on our collaboration, for all his advice and help on academic and all the other matters, and for all the things I learned from him during my Ph.D. study.

I am particularly thankful to Shang-Hua Teng for providing me generous support and sharing his insightful thoughts with me. I also thank László Babai for giving me lots of suggestions, and for guiding my study on graph isomorphism problem.

Thanks to Rocco Servedio and Mihalis Yannakakis for the wonderful collaboration and serving on my thesis committee.

Thanks to my collaborators over the past few years: Zeyuan Allen-Zhu, Wei Chen, Siu-on Chan, Alex Collins, Rachel Cummings, Ilias Diakonikolas, Ming Duan, Te Ke, Xuejia Lai, Chin-Yew Lin, Zhenming Liu, Pinyan Lu, Henry Lam, Michael Mitzenmacher, Ryan O'Donnell, Anthi Orfanou, Dimitris Paparas, David Rincon, Tao Sun, Li-Yang Tan, Bo Tang, Yajun Wang, Wei Wei, John Wright, Mohan Yang, Yifei Yuan, Ming Zhang, Yu Zhao, Bo Zhu. A special thanks to John Wilmes and Krzysztof Onak for the great effort on contributions to the joint works as part of this thesis.

Finally, many thanks to my parents Wangong and Yufen for their constant support and encourangement. Thanks to my wife Wei and my son Pingyue, for their love and accompany over the years.

To my parents.

Chapter 1

Introduction

One of the most fascinating graph-theoretic problems is to determine whether two graphs are isomorphic to each other [Read and Corneil, 1977]. In this problem, we are given two graphs $G = (V, E)$ and $H = (V, E')$ on the same set of vertices, and are asked to decide whether there exists a permutation σ such that for all pairs of vertices (u, v) in V, $(u, v) \in E$ if and only if $(\sigma(u), \sigma(v)) \in E'$.

It follows from the theory of interactive proofs that the Graph Isomorphism problem (GI) is not NP-complete unless the polynomial-time hierarchy collapses ([Goldreich *et al.*, 1991; Babai, 1985; Goldwasser *et al.*, 1985; Boppana *et al.*, 1987; Goldwasser and Sipser, 1986], see [Babai and Moran, 1988] for a self-contained proof). On the other hand, polynomial-time algorithms have been developed for special families of graphs. It was also shown in [Babai and Kucera, 1979; Babai *et al.*, 1980; Czajka and Pandurangan, 2008] that isomorphism testing for random graphs is easy. For general graphs, the previous best upper bound was $\exp(\widetilde{O}(\sqrt{n}))$ where n is the number of vertices and the tilde hides polylog factors [Babai and Luks, 1983; Babai *et al.*, 1983; Zemlyachenko *et al.*, 1982]. This upper bound was significantly improved to $\exp((\log n)^{O(1)})$ by Babai [Babai, 2015] recently.

The families of highly regular structures have been perceived as difficult cases for graph isomorphism. These highly regular structures arise naturally as obstacles for both the classical group theory and combinatorial approaches for the graph isomorphism problem [Babai, 2014].

In this thesis, we develop new structure theory for the highly regular structures, includ-

ing Steiner 2-designs, strongly regular graphs and primitive coherent configurations.

For all these highly regular structures, we show some new bounds on the rate of expansion of small sets of vertices in certain ranges of parameters. We also prove the existence of clique geometries for strongly regular graphs and primitive coherent configurations in other cases. A clique geometry of a graph is a collection of maximal cliques such that every edge belongs to a unique clique. Clique geometries allows us to separate the exceptions with large automorphism groups from the others with nice claw structures.

Based on these new structure theory of the highly regular structures, we study the isomorphism testing, as well as the number of automorphisms for the highly regular structures. The latter is of interest to algebraic combinatorics. We use the new structure results to prove the efficiency of the *individualization/refinement* procedure on the highly regular structures.

One classical combinatorial approach for graph isomorphism is the naive refinement method: one first assigns each vertex a label that is equal to its degree and then repeatedly relabels each vertex based on the set of labels of its neighbors. It is easy to show that, if vertices of a graph G eventually obtain distinct labels under this process, then the testing of isomorphism involving G with any other graph can be solved in polynomial time, and G has only trivial automorphism. However, this method fails to distinguish any pair of vertices for regular graphs. To break this symmetry of regular graphs, a more powerful technique is the individualization/refinement procedure. One first chooses a set of (a small number of) vertices, which we will refer to as the seeding set, and assigns each vertex in it a distinct label to jump start the refinement procedure. It is standard that if G has a seeding set S of size k such that the individualization of S followed by refinement assigns a distinct label to every vertex of G, then the test of isomorphism involving G with any other graph can be solved in time $n^k \cdot \text{poly}(n)$, and G has at most n^k automorphisms.

1.1 Summary of thesis contributions

We present the following results in this thesis.

1. We show that the individualization of $O(\log v)$ points and lines suffices for refine-

ment to completely split a Steiner 2-design, where v is the number of points in the Steiner 2-design. A Steiner 2-design consists of points and lines, where each line passes through the same number of points and each pair of points uniquely determines a line. Each Steiner 2-design induces a Steiner graph, in which vertices represent lines and edges represent intersections of lines. Steiner graphs are an important subfamily of strongly regular graphs whose isomorphism testing has challenged researchers for years. Our result implies a quasipolynomial-time algorithm for isomorphism testing of Steiner 2-designs. This improves the previous best bound of $\exp(\widetilde{O}(v^{1/4}))$ by Spielman [Spielman, 1996]. Before our result, quasipolynomial-time isomorphism testing was only known for the case when the line size is polylogarithmic, as shown by Babai and Luks [Babai and Luks, 1983]. Our result also implies that every Steiner 2-design has at most quasipolynomial automorphisms.

2. We present an $\exp(\widetilde{O}(n^{1/5}))$-time algorithm for isomorphism testing of strongly regular graphs, improving the best previous $\exp(\widetilde{O}(n^{1/3}))$ bound by Spielman [Spielman, 1996]. A strongly regular graph is a regular graph such that for each pair of vertices, the number of their common neighbors is determined solely by whether they are connected.

   The previous results on isomorphism testing of strongly regular graphs [Babai, 1980; Spielman, 1996] were based on the analysis of the classical individualization/refinement method. Our new bound is based on a combination of a deeper analysis of the individualization/refinement method with Luks's group theoretic divide-and-conquer methods [Luks, 1982].

   Following Spielman's work [Spielman, 1996], our analysis builds on Neumaier's 1979 classification of strongly regular graphs [Neumaier, 1979]. One of Neumaier's classes, the aforementioned Steiner graphs, has been eliminated as a bottleneck by showing a quasipolynomial time algorithm for the isomorphism testing of Steiner 2-designs. In the remaining hard cases, we have the benefit of "Neumaier's claw bound" and its asymptotic consequences derived by Spielman.

   We also prove the following purely combinatorial result: Any non-trivial and non-

graphic strongly regular graph has a set of $\widetilde{O}(n^{9/37})$ vertices whose individualization and refinement completely split the graph. This implies that the order of the automorphism group of non-trivial and non-graphic strongly regular graphs is at most $\exp(\widetilde{O}(n^{9/37}))$, improving an earlier $\exp(\widetilde{O}(n^{1/3}))$ bound by Spielman [Spielman, 1996].

3. We show that after excluding easily described and recognized exceptions, every primitive coherent configuration has a set of $\widetilde{O}(n^{1/3})$ vertices whose individualization and subsequent refinement completely split the configuration, improving the best previous $\exp(\widetilde{O}(n^{1/2}))$ bound by Babai [Babai, 1981b]. Primitive coherent configurations are colored directed graphs that generalize strongly regular graphs. Moreover, primitive coherent configurations arise naturally as obstacles to combinatorial divide-and-conquer approaches for general graph isomorphism. In a natural sense, the isomorphism problem for primitive coherent configurations is a stepping stone between strongly regular graph isomorphism and the general graph isomorphism problem.

   The emergence of exceptions illuminates the technical difficulties: we had to separate these cases from the rest. For the analysis we develop a new combinatorial structure theory for PCCs that in particular demonstrates the presence of "asymptotically uniform clique geometries" among the constituent graphs of PCCs in a certain range of the parameters. Our result also implies an $\exp(\widetilde{O}(n^{1/3}))$-time algorithm for isomorphism testing of primitive coherent configurations and an $\exp(\widetilde{O}(n^{1/3}))$ upper bound on the number of automorphisms of PCCs (with known exceptions), making the first progress in 33 years on an old conjecture of Babai (If a PCC has at least $\exp(n^{\varepsilon})$ automorphisms for some positive constant ε, then the automorphism group of the PCC is a primitive permutation group).

   A corollary to our result is an $\exp(\widetilde{O}(n^{1/3}))$ upper bound on the order of primitive but not doubly transitive permutation groups (with known exceptions). This bound. This bound was previously known only through the Classification of Finite Simple Groups [Cameron, 1981].

   The complexity of testing isomorphism of PCCs has recently been improved to quasipolynomial time by Babai's general graph isomorphism algorithm [Babai, 2015]. Our

structural results and automorphism bounds for PCCs are not affected by Babai's new result.

In addition, PCCs play a prominent role in Babai's new algorithm for general graph isomorphism. Further progress on the old conjecture of Babai has the potential of simplifying Babai's new algorithm using a deeper combinatorial analysis.

4. We also investigate the isomorphism testing problem for two unknown graphs in the property testing setting. In this setting, we want to distinguish pairs of graphs that are isomorphic from pairs of graphs that are significantly different. We say that two graphs are ε-far if at least $\varepsilon\binom{n}{2}$ edges must be added or removed from one graph in order to make the two graphs isomorphic. The goal of the property testing algorithm is to accept with probability at least 9/10 if the input graphs are isomorphic and reject with probability at least 9/10 if the input graphs are ε-far, for some given constant $\varepsilon > 0$. We study the question of how many queries are required to distinguish between the two cases. A query is defined as asking if a two vertices are adjacent or not.

   We present a new property testing algorithm using $n \cdot 2^{O(\sqrt{\log n})}$ samples with an $\exp(2^{O(\sqrt{\log n})})$ running time, improving previous $\widetilde{O}(n^{5/4})$ sample complexity by Fischer and Matsliah [Fischer and Matsliah, 2008]. This new sample complexity matches the information theoretical lower bound up to some multiplicative subpolynomial factor.

1.2 Acknowledgement of collaborations

The ingredients of this thesis are based on joint works with László Babai, Xi Chen, Krzysztof Onak, Shang-Hua Teng and John Wilmes. We acknowledge the papers where these joint results have appeared or will appear in this section. A simultaneous Ph.D. thesis by Wilmes [Wilmes, 2016] will include disjoint elements from the related joint works.

The isomorphism of Steiner 2-designs is a joint work with Xi Chen and Shang-Hua Teng, and was published in the proceeding of 45th ACM Symposium on Theory of Computing [Chen *et al.*, 2013]. A result essentially identical to ours obtained simultaneously by Babai and Wilmes was published in the same proceeding [Babai and Wilmes, 2013], and

included in [Wilmes, 2016].

The isomorphism of strongly regular graphs is a joint work with László Babai, Xi Chen, Shang-Hua Teng and John Wilmes, and was published in the proceeding of 54th Annual IEEE Symposium on Foundations of Computer Science [Babai *et al.*, 2013]. The technical details presented in this thesis appears in the journal version of the same paper. Other elements of this work are included in Wilmes's Ph.D. thesis [Wilmes, 2016].

The automorphism bound of strongly regular graphs is based on an unpublished joint work with Xi Chen and Shang-Hua Teng.

The isomorphism of primitive coherent configuration is a joint work with John Wilmes, and was published in the proceeding of 47th ACM Symposium on Theory of Computing [Sun and Wilmes, 2015]. The technical details presented in this thesis appear in the journal version of the same paper. Other elements of the proof of this work are included in Wilmes's Ph.D. thesis [Wilmes, 2016].

The property testing of graph isomorphism is based on an unpublished joint work with Krzysztof Onak.

1.3 Organization of the thesis

The following chapters are organized as follows: We begin in Chapter 2 by giving some basic definitions, introducing the individualization and refinement method. We present our quasi-polynomial time algorithm for isomorphism of Steiner 2-designs in Chapter 3. In Chapter 4, we show our $\exp(\widetilde{O}(n^{1/5}))$ time algorithm for isomorphism of strongly regular graphs, and the $\exp(\widetilde{O}(n^{9/37}))$ upper bound on the number of automorphisms. In Chapter 5, we present the new result for primitive coherent configurations. In Chapter 6, we present our result on the property testing of graph isomorphism.

Chapter 2

Preliminaries

2.1 Graph isomorphism and related problems

We formally define the graph isomorphism problem.

Definition 2.1.1. *Let $G = (V, E)$ and $G = (V', E')$ be two graphs. We say a bijection $\sigma : V \to V'$ is an isomorphism mapping from G to G' if for any pair of vertices $u, v \in V$, $(\sigma(u), \sigma(v)) \in E'$ iff $(u, v) \in E$.*

Problem 1. (Graph isomorphism problem) *Given two graphs $G = (V, E)$ and $G' = (V', E')$, the graph isomorphism problem is to decide whether there is an isomorphism bijection between the two graphs.*

Graph isomorphism problem is closely related to the graph canonical form problem.

Definition 2.1.2. *Let $\mathcal{K}$ denote a family of graphs that is closed under isomorphism. A map $F : \mathcal{K} \to \mathcal{K}$ is called a* canonical form *for graphs in $\mathcal{K}$ if*

1. *For any G in $\mathcal{K}$, $F(G)$ and G are isomorphic;*
2. *For any $G, H \in \mathcal{K}$, if G and H are isomorphic, then $F(G) = F(H)$.*

Problem 2. (Graph canonical form problem) *Let $\mathcal{K}$ be a family of graphs that is closed under isomorphism. The canonical form problem is to define a function $F : \mathcal{K} \to \mathcal{K}$ such that*

1. *F satisfies Definition 2.1.2;*

2. *F is computable in within a specific time bound $f(n)$.*

For instance, if we set $f(n)$ to be $n! \cdot \text{poly}(n)$, then a corresponding F can be defined as the graph in $\mathcal{K}$ with lexicographically smallest adjacency matrix.

It is straightforward that the graph isomorphism problem is not harder than computing a canonical form: Isomorphism of members of $\mathcal{K}$ can be decided by two applications of a canonical form function and comparison of the outputs.

An automorphism of graph G is an isomorphism from G to G. One interesting question in algebraic combinatorics is to upper bound the number of automorphisms of highly regular structures. We contribute three such upper bounds: a bound for Steiner 2- designs, one for strongly regular graphs, and one for PCCs (with known exceptions). The bound for PCCs implies the same bound on the order of primitive permutation groups; this bound was previously only known through the CFSG.

All the above definitions for graphs can be naturally extended to edge colored graphs and finite geometries.

2.2 Individualization and refinement

A classical heuristic to GI is the *individualization/refinement* (I/R) method. *Individualization* means the assignment of individual colors to some vertices; then the irregularity so created propagates via some canonical color refinement process. For a class $\mathcal{C}$ of graphs, an assignment $G \mapsto G'$ is a *color refinement* if $G, G' \in \mathcal{C}$ have the same set of vertices and the coloring of G' is a refinement of the coloring of G. Such an assignment is *canonical* if for all $G, H \in \mathcal{C}$, we have $\text{Iso}(G, H) = \text{Iso}(G', H')$, where $\text{Iso}(G, H)$ denotes the set of isomorphisms from G to H. In particular, $\text{Aut}(G) = \text{Aut}(G')$, where $\text{Aut}(G)$ denotes the automorphisms of G.

The simplest canonical color refinement process is the *naive vertex refinement*. The edge-colors do not change, only the vertex-colors are refined. The refined color of vertex u of the graph G encodes the following information: the current color of u and the number of vertices v of color i adjacent to u, for every vertex-color i. We note that it can be performed

in polynomial time. Repeated application of the naive vertex refinement process leads to the *stable refinement* after at most $n-1$ rounds.

In 1968, Weisfeiler and Leman defined another natural canonical refinement process of colorings of the ordered pairs [Weisfeiler and Leman, 1968]. This refinement was generalized to k-dimensional Weisfeiler-Leman refinement. The k-dimensional Weisfeiler-Leman refinement initially assigns colors to all the ordered k-tuples of the vertices according to the canonical form of the induced subgraph of the k vertices. At each step of the refinement, the color of every k-tuple is further updated by considering the ordered multiset of colors of the neighbors of the given k-tuple (here the neighbors are the k-tuples differing in exactly one element). The vertex color of Weisfeiler-Leman refinement for a given vertex is the color of the tuple with k copies of the given vertex. We note that every round of k-dimensional Weisfeiler-Leman refinement can be performed in time $n^{k+O(1)}$, and the refinement is stable after at most n^k rounds.

If after individualizing the elements of a set $S \subseteq V$, all vertices get different colors in the resulting stable refinement, we say that S *completely splits* G (with respect to the given canonical refinement process).

The following lemma is standard (see, e.g., [Babai *et al.*, 2013, Section 2]).

Lemma 2.2.1. *Let $\mathcal{K}$ be a class of graphs, and suppose that for every $G \in \mathcal{K}$ there is set of α vertices that completely splits G with respect to a polynomial-time canonical color refinement process. Then the following statements hold for every $G \in \mathcal{K}$:*

1. *$|\operatorname{Aut}(G)| \leq n^{\alpha}$;*
2. *a canonical form for G can be computed in time $n^{\alpha+O(1)}$;*
3. *for every $H \in \mathcal{K}$, the set of isomorphisms from G to H can be listed in time $n^{\alpha+O(1)}$.*

In particular, I/R can efficiently split a graph G only if G has a small automorphism group.

2.2.1 Strong labeling

Without loss of generality, in this subsection we assume that $G = (V, E)$ is a graph on n vertices and $V = [n] = \{1, \ldots, n\}$.

Let $\mathcal{K}$ denote a family of graphs with trivial automorphism group that is closed under isomorphism. Canonical forms of graphs in $\mathcal{K}$ can be obtained from a strong labeling L which takes two parameters $G = (V, E)$ and $v \in V$, and returns a binary string $L(G, v)$:

Definition 2.2.2. (Strong Labeling) *Let $\mathcal{K}$ denote a family of graphs with trivial automorphism group that is closed under isomorphism. A labeling L is* strong *for graphs in $\mathcal{K}$ if for every graph $G \in \mathcal{K}$, L satisfies the following two properties:*

- **Invariant Under Isomorphism:** *For every $v \in V$ and $\sigma \in \mathsf{Sym}(V)$, $L(G, v) = L(\sigma(G), \sigma(v))$, where we use $\sigma(G)$ to denote the graph $G^* = (V, E^*)$ such that $(u, v) \in E$ if and only if $(\sigma(u), \sigma(v)) \in E^*$.*
- **Distinctness:** *For every two distinct vertices $u, v \in V$, we have $L(G, u) \neq L(G, v)$.*

A canonical form F for $\mathcal{K}$ can be derived from a strong labeling L for $\mathcal{K}$ as follows. Given $G = (V, E)$ in $\mathcal{K}$ with $V = [n]$, set $F(G) = \sigma(G)$, where $\sigma \in \mathsf{Sym}(V)$ and $\sigma(u) \in [n]$ is set to be the rank of $L(G, u)$ among $\{L(G, v) : v \in V\}$, under the lexicographical order.

With Definition 2.2.2, we review the process of individualization and refinement as follows (with naive vertex refinement as an example):

- *Individualization*: We first select a set of t vertices from $G = (V, E)$ and assign each vertex in it a special and distinct label. So for convenience we use a (not necessarily injective) map $f : [t] \to V$, called a *seeding map*, to specify a subset of V such that $f(i)$ is assigned the ith special label, $i \in [t]$. We may also use a feature selection algorithm to assign each vertex outside of $f[t] = \{f(i) : i \in [t]\}$ an initial label.
- *Naive vertex refinement*: At each step, each vertex in $V - f[t]$ receives the multiset whose elements are the multisets of labels of its neighbors (including its neighbors in $f[t]$) as well as itself, and obtains a new label that is set to its rank among all multisets of vertices in $V - f[t]$, under the lexicographical order. It continues this refinement process until no more progress can be made. Note that vertices in $f[t]$ always keep their initially assigned special labels. It is clear that this process terminates in polynomial time.

We call $f : [t] \to V$ a *good* seeding map of parameter t for G, if the individualization of f followed by the refinement process produces a distinct labeling of vertices of G. Suppose $H = (V, E')$ is isomorphic to G, and let σ denote an isomorphism from G to H. It is easy to show that $\sigma \circ f$ is also a good seeding map for H. Moreover, the final label of each vertex u in G must be the same as that of $\sigma(u)$ in H after refinement, as long as f and $\sigma \circ f$ are individualized, respectively, using the same set of matching special labels. As a result, if G is guaranteed to have a good seeding map of parameter t, then one can test isomorphism involving G in time $n^t \cdot \text{poly}(n)$, since there are $O(n^t)$ many seeding maps of parameter t, and for each of them the refinement process terminates in time polynomial in n.

In general, the individualization and refinement method can be viewed as a strong labeling L that takes three parameters: a graph $G = (V, E)$, a seeding map $f : [t] \to V$, and a vertex $u \in V$; and returns a binary string. It fits the following definition:

Definition 2.2.3. *Let $\mathcal{K}$ denote a family of graphs that is closed under isomorphism, and let $T : \mathbb{N} \to \mathbb{N}$ denote an integer function. We say a labeling L is* T-strong *for graphs in $\mathcal{K}$ if for all $G = (V, E) \in \mathcal{K}$ on n vertices, there exists a* good *seeding map $f : [t] \to V$ with $t = T(n)$ satisfying the following two properties:*

- **Invariant Under Isomorphism:** *For all $u \in V$, and $\sigma \in \mathsf{Sym}(V)$, we have $L(G, f, u) = L(\sigma(G), \sigma \circ f, \sigma(u))$.*
- **Distinctness:** *For every two distinct vertices $u, v \in V$, we have $L(G, f, u) \neq L(G, f, v)$.*

We can also derive a canonical form F for $\mathcal{K}$ from a T-strong labeling L for $\mathcal{K}$ [Babai, 1980]. Given G we enumerate all seeding maps $f : [t] \to V$, where $t = T(|V|)$, and keep only the good ones. For each good f, let $\sigma_f \in \mathsf{Sym}(V)$ denote the permutation in which $\sigma_f(u)$ is the rank of $L(G, f, u)$ among $\{L(G, f, v) : v \in V\}$, under the lexicographical order. Finally we set $F(G) = H$, where H has the lexicographically smallest adjacency matrix among $\{\sigma_f(G)\}$. Note that the time needed to compute F depends exponentially on $T(n)$.

2.2.2 Canonical pairwise distinguisher

One way to produce the strong labeling is to construct pairwise distinguisher.

We use two procedures named **Cert** and **Test** to show that after individualizing some vertices every graph completely splits. The input variables of **Cert** and **Test** are:

1. The four input variables of **Cert** include a graph $G = (V, E)$, a vertex-seeding map $f : [t] \to V$ for some integer $t \geq 1$, and two distinct vertices $x, y \in V$.

2. The four input variables of **Test** are the same, except that M is a binary string called a *certificate* (see below) and there is only one vertex $x \in V$.

Note that f is not necessarily an injective map.

The output of **Cert** is either nil, in which case we say it fails, or a binary string M that encodes a certificate. The output of **Test**, on the other hand, is either 0 or 1.

Let $M_{x,y}$ denote the output of $\mathbf{Cert}(G, f, x, y)$. We use $M_{x,y}$ to encode some information that can be used to distinguish x and y. And $M_{x,y}$ can be realized by procedure **Test**: If the first three input variables of **Test** are $G, f, M_{x,y}$, then **Test** outputs 1 when the fourth input variable is x, and outputs 0 when the fourth input variable is y. In addition, we also require that the two procedures give the same output if distinct inputs are actually identical under isomorphism.

Formally we will refer to a pair of procedures that satisfy the property below as a *canonical pairwise distinguisher*

Property 2.2.4. (Canonical Pairwise Distinguisher) **Cert** *and* **Test** *are two procedures that satisfy:*

- **Invariant Under Isomorphism***: Let $G = (V, E)$ and $G' = (V', E')$ denote two isomorphic graphs, and ϕ denote an isomorphism from G to G'. For all pairs of vertices $x, y \in V$, $t \geq 1$, all seeding map $f : [t] \to V$, and for all binary strings M, we have*

$$\mathbf{Cert}\,(G, f, x, y) = \mathbf{Cert}\left(G', \phi \circ f, \phi(x), \phi(y)\right) \quad \text{and}$$
$$\mathbf{Test}\,(G, f, M, x) = \mathbf{Test}\left(G', \phi \circ f, M, \phi(x)\right)$$

where $\phi \circ f$ denotes the seeding map from $[t]$ to V' with $\phi \circ f(i) = \phi(f(i)) = \{\phi(x) : x \in f(i)\}$.

- **Pairwise Distinctness**: *If* $M = \mathbf{Cert}(G, f, x, y) \neq \text{nil}$ *for some* t *and* f*, then we have*

$$\mathbf{Test}(G, f, M, x) \neq \mathbf{Test}(G, f, M, y).$$

The pairwise-distinctness condition does not impose any condition over $\mathbf{Test}(\mathcal{S}, f, M, x)$ and $\mathbf{Test}(\mathcal{S}, f, M, y)$, if M is not the certificate output by $\mathbf{Cert}(\mathcal{S}, f, x, y)$.

We show that a canonical pairwise distinguisher for G can be used to derive a canonical form for all graphs isomorphic to G, if there exists a seeding map f of parameter t such that

$$\mathbf{Cert}(G, f, x, y) \neq \text{nil}, \quad \text{for all pairs of distinct vertices } x, y \in V.$$

To this end, given any $H = (V, E')$ isomorphic to G (including G itself), we enumerate all possible seeding map f' of parameter t, and only keep those satisfy

$$\mathbf{Cert}\left(H, f', x, y\right) \neq \text{nil}, \quad \text{for all pairs of distinct vertices } x, y \in V.$$

By our assumption we know such seeding map exists. For each such a seeding map f', we let

$$M_{x,y} = \mathbf{Cert}\left(H, f', x, y\right) \neq \text{nil}$$

and use M_{sort} to denote the vector of $m = \binom{n}{2}$ entries obtained by sorting all the m certificates $M_{u,v}$ according to the lexicographical order. Then we let $\sigma_{H,f'}$ denote the following permutation: $\sigma_{H,f'}(x)$ is set to be the rank of the following length-m binary string associated with x:

$$\left(\mathbf{Test}\left(H, f', M_{\text{sort}}(1), x\right), \ldots, \mathbf{Test}\left(H, f', M_{\text{sort}}(m), x\right)\right)$$

among the n strings associated with vertices in V under the lexicographical order. (Here by the property of pairwise distinctness we know that all the n strings associated with vertices in V are distinct.) Finally we set $F(H)$ to be the graph with the lexicographically smallest adjacency matrix among $\{\sigma_{H,f'}(H)\}$. The following lemma shows that F is a canonical form: $F(H) = F(G)$, for all H isomorphic to G.

Lemma 2.2.5. *Let* $(\mathbf{Cert}, \mathbf{Test})$ *denote a canonical pairwise distinguisher for* $G = (V, E)$ *with* $V = [n]$*. Let* $H = \sigma(G)$ *be a graph isomorphic to* G*. Suppose for some positive*

integer t, $f : [t] \to V$ is a map such that for all pairs of vertices $x, y \in V$,
$M_{x,y} = \mathbf{Cert}\,(G, f, x, y) \neq nil$. Then we have

$$\sigma_{G,f}(G) = \sigma_{H,f'}(H), \qquad \textit{where } f' = \sigma \circ f.$$

Proof. First of all, because **Cert** is invariant under isomorphism, we have

$$\mathbf{Cert}\,(G, f, x, y) = \mathbf{Cert}\,\big(H, f', \sigma(x), \sigma(y)\big)$$

This implies the two tuples M_{sort} and M'_{sort} of m certificates constructed from G, f and H, f', respectively, are exactly the same. Then because **Test** is invariant under isomorphism, we have

$$\mathbf{Test}\,(G, f, M_{\text{sort}}(i), x) = \mathbf{Test}\,\big(H, f', M_{\text{sort}}(i), \sigma(x)\big)$$

and thus, the two strings associated with x in G and $\sigma(x)$ in H are exactly the same. It follows that

$$\sigma_{G,f}(x) = \sigma_{H,f'}(\sigma(x))$$

As a result, we have $\sigma_{G,f}(G) = \sigma_{H,f'}(H)$ and the lemma follows. □

Corollary 2.2.6. *Let $\mathcal{K}$ be a class of graphs. If for every graph $G \in \mathcal{K}$, there exists a seeding map t vertices and a canonical pairwise distinguisher, then*

1. $|\operatorname{Aut}(G)| \leq n^t$;
2. *If the canonical pairwise distinguisher can be computed in time α, then the canonical form of G can be computed in time $n^{t+O(1)} \cdot \alpha$.*

In this thesis, all pairwise distinguishers presented are superseded by the classical Weisfeiler-Leman refinement, which means that if there is a pairwise distinguisher for a graph with a seeding map, then the graph is completely split by the classical WL refinement after individualizing the seeds.

Chapter 3

Isomorphism of Steiner 2-designs

In this chapter we analyze the structure and isomorphism of Steiner 2-designs. We show that every Steiner 2-design is completely split by a set of vertices of logarithmic size

3.1 Steiner 2-designs

Definition 3.1.1. (STEINER 2-DESIGNS) *A* Steiner 2-design *with parameters* (v, n, s, h) *is a pair* $\mathcal{S} = (\mathcal{P}, \mathcal{L})$ *that satisfies the following conditions: 1)* $\mathcal{P}$ *is a set of* $v = |\mathcal{P}|$ *points; 2)* $\mathcal{L}$ *is a set of* n *lines, where each line* $L \in \mathcal{L}$ *is a subset of* $\mathcal{P}$ *of cardinality* $|L| = s$*; 3) For any two distinct points* $p, q \in \mathcal{P}$*, there exists a unique line* $L \in \mathcal{L}$ *such that* $p, q \in L$*; and 4) Each point* $p \in \mathcal{P}$ *belongs to exactly* h *lines.*

Each Steiner 2-design $\mathcal{S}$ induces a Steiner graph G as follows: vertices of G correspond to lines of $\mathcal{S}$ and two vertices are adjacent in G if and only if their corresponding lines intersect in $\mathcal{S}$. It is worth mentioning that each point of $\mathcal{S}$ corresponds to a clique of size h in G.

We have the following basic property of Steiner 2-designs:

Proposition 3.1.2 (Basic)**.** *The parameters* (v, n, s, h) *of a Steiner* 2*-design must satisfy* $vh = ns$ *and*

$$\binom{v}{2} = n\binom{s}{2}.$$

Moreover, from $ns(s-1) = v(v-1)$ and $v \geq s \geq 2$, we have

$$\frac{v}{s} \leq h = \frac{ns}{v} = \frac{v-1}{s-1} < \frac{v}{s} \cdot \frac{s}{s-1} \leq 2 \cdot \frac{v}{s}. \tag{3.1}$$

Using (3.1), we also get the following useful inequalities:

$$\sqrt{\frac{n}{2}} \leq \frac{v}{s} \leq \sqrt{n} \quad \text{and} \quad \sqrt{n} \leq h = \frac{ns}{v} \leq \sqrt{2n} \tag{3.2}$$

3.2 Isomorphism of Steiner 2-designs

We focus on the isomorphism testing problem of Steiner 2-designs:

Definition 3.2.1. (Isomorphisms between Steiner 2-designs). *Let $\mathcal{S} = (\mathcal{P}, \mathcal{L})$ and $\mathcal{S}' = (\mathcal{P}, \mathcal{L}')$ denote two Steiner 2-designs on the same set of points $\mathcal{P}$. We say $\phi \in \mathsf{Sym}(\mathcal{P})$ is an* isomorphism *from $\mathcal{S}$ to $\mathcal{S}'$ if it induces a bijection from $\mathcal{L}$ to $\mathcal{L}'$: $L \in \mathcal{L}$ if and only if $\phi(L) \in \mathcal{L}'$, where $\phi(L) = \{\phi(p) : p \in L\}$.*

Before this and the work of Babai and Wilmes [Babai and Wilmes, 2013], Spielman's $\exp(\widetilde{O}(n^{1/4}))$ time bound [Spielman, 1996] was the best bound on the complexity of testing isomorphism of Steiner 2-designs. For the special case when the line size is 3, Miller obtained an $(n^{\log n + O(1)})$-time algorithm in [Miller, 1978]. Later, Babai and Luks gave a quasipolynomial-time algorithm for isomorphism of Steiner 2-designs of polylogarithmic line size [Babai and Luks, 1983].

In this chapter, we give an $n^{O(\log n)}$-time isomorphism-testing algorithm for general Steiner 2-designs. Our approach is inspired by the individualization/refinement method and the previous analyses of Babai [Babai, 1981a] and Spielman [Spielman, 1996] over strongly regular graphs. The proofs of [Babai, 1981a; Spielman, 1996] focus on showing that a small set of randomly chosen vertices (seeding set) suffices to *distinguish* each pair u and v of vertices with high probability, i.e., refinement after individualizing the seeding set assigns distinct labels to u and v. Then the existence of a small seeding set whose individualization results in a distinct colors of all vertices follows by a union bound.

In order to distinguish a pair of vertices u, v, Babai and Spielman examine structures rooted at u and v, respectively, and show that they interact with the seeding set differently, with high probability. Their structures rooted at u and v are closely related to the refinement

process so that having different interactions with the seeding set directly implies that refinement (in one [Babai, 1981a] or two [Spielman, 1996] steps) assigns distinct labels to u and v.

Influenced by [Babai, 1981a; Spielman, 1996], we consider an isomorphism-testing framework for Steiner 2-designs. It uses a small number of random seeding points and lines to build multi-stage combinatorial structures to distinguish each pair p, q of points with high probability. By distinguishing a pair of points, we again mean that the multi-stage structures built from p and q, respectively, interact with the seeding set differently. Note that, for the purpose of isomorphism testing, these structures do not need to be tightly coupled with the standard refinement process provided they satisfy the isomorphism-invariant condition, meaning (informally) that mapping everything (p, q and the seeding set) to an isomorphic Steiner 2-design would result in exactly the same structures and interactions. Thus, while our multi-stage structures are designed with intention to capture the multi-step label propagation of the refinement process, we use this relaxation to fine-tune the structures and gain better control of their analysis, which coincidentally leads us to deviate from the standard refinement process (see more discussion below).

The main question is then: How to design isomorphism-invariant structures so that a small random seeding set suffices to distinguish each pair of points with high probability? To this end, we give a construction of multi-stage structures for which a seeding set of size $O(\log v)$ suffices.

Theorem 3.2.2. *Every Steiner 2-design of v points is completely split by $O(\log v)$ individualizations under classical Weisfeiler-Leman refinement.*

This leads to a quasipolynomial-time algorithm to compute a canonical form for Steiner 2-designs (and thus isomorphism-testing) :

Corollary 3.2.3. *A canonical form for Steiner 2-designs with v vertices can be found in time $v^{O(\log v)}$. As a consequence, isomorphism of Steiner 2-designs can be decided, and the set of isomorphisms found, within the same time bound.*

And it also implies an upper bound of $v^{O(\log v)}$ on the number of automorphisms of nontrivial Steiner 2-designs.

Corollary 3.2.4. *Every Steiner 2-designs with v vertices has at most $v^{O(\log v)}$ automorphisms.*

The best previous bound was $\exp(\widetilde{O}(\sqrt{v}))$ by Babai-Pyber [Babai and Pyber, 1994] and Spielman [Spielman, 1996].

3.3 Multi-stage design

In this section, we give a high-level description of our multi-stage structures designed to distinguish all pairs of points. Given a Steiner 2-design $\mathcal{S} = (\mathcal{P}, \mathcal{L})$, we pick a point seeding map f by simply drawing t points from $\mathcal{P}$ with replacement, independently and uniformly at random. We also pick a line seeding map g similarly.

Fix an arbitrary pair p, q of points. Our goal is then to build two isomorphism-invariant structures $\Pi(p)$ for p and $\Pi(q)$ for q, respectively, so that they interact differently with the seeding maps with high probability, and p, q are distinguished from each other. As described earlier, $\Pi(p)$ and $\Pi(q)$ are designed to mimic label propagation in the refinement process. Roughly speaking, the structure $\Pi(p)$ built for p is a tree rooted at p (level 1) in which points at level i are those that may affect the label of p after $2(i-1)$ steps, in some way, by propagating along a sequence of $i-1$ lines.

In this context, for example, Babai uses the seed vertices to interact with $\Pi(u)$, a single-level structure with u at the root [Babai, 1981b]; Spielman [Spielman, 1996] uses the seed vertices to interact with $\Pi(u)$, a two-level structure with u at the root and $N(u)$ at the second level, where $N(u)$ denotes the set of k neighbors of u.

Our construction of $\Pi(p), \Pi(q)$, unlike [Babai, 1981b] and [Spielman, 1996], is assisted by a subset of points and lines from the seeding maps. To build a τ-level structure with seeds $R = f[t] \cup g[t]$, we first partition R into τ disjoint sets $R_1, \ldots, R_\tau$. We use elements from $R_1, \ldots, R_{\tau-1}$ to iteratively construct the two multi-level structures $\Pi(p)$ and $\Pi(q)$, level by level, and at the end use R_τ to interact with the last level of $\Pi(p)$ and $\Pi(q)$.

The starting levels of $\Pi(p), \Pi(q)$ are $\Pi_1(p) = (\{p\})$ and $\Pi_1(q) = (\{q\})$, respectively. We use elements from R_1 to define from $(\Pi_1(p), \Pi_1(q))$ a *tuple of subsets of* $\mathcal{P}$ to form the second level $\Pi_2(p)$ of $\Pi(p)$, and a tuple of subsets of $\mathcal{P}$ to form $\Pi_2(q)$. So, level by level, we

define $(\Pi_\ell(p), \Pi_\ell(q))$ using $(\Pi_{\ell-1}(p), \Pi_{\ell-1}(q))$ and $R_{\ell-1}$, by exploiting interactions of their neighborhood structures.

In our construction, $\Pi_\ell(p), \Pi_\ell(q)$ are each a tuple of subsets of $\mathcal{P}$, with the same length denoted by m_ℓ. Let

$$\Pi_\ell(p) = \big(A_{\ell,i}\big)_{i\in[m_\ell]} \quad \text{and} \quad \Pi_\ell(q) = \big(B_{\ell,i}\big)_{i\in[m_\ell]}$$

for some $m_\ell \geq 1$. Our construction ensures that for each $i \in [m_\ell]$, if set $A_{\ell,i}$ is defined by the interaction between a seed $r \in R_{\ell-1}$ and a set $A_{\ell-1,j} \in \Pi_{\ell-1}(p)$, for some j in $[m_{\ell-1}]$, then $B_{\ell,i} \in \Pi_\ell(q)$ is defined by the interaction between the same $r \in R_{\ell-1}$ and $B_{\ell-1,j} \in \Pi_{\ell-1}(q)$.

In other words, $\Pi(p)$ and $\Pi(q)$ are two isomorphic branching structures of τ levels, such that

- Each path from the root $\Pi_1(p)$ (or $\Pi_1(q)$) to a leaf set in $\Pi_\tau(p)$ (or $\Pi_\tau(q)$) is associated with a unique seed sequence from $R_1 \times R_2 \times \cdots \times R_{\tau-1}$; and
- For each $i \in [m_\tau]$, the associated seed sequence of the path from root $\Pi_1(p)$ to $A_{\tau,i}$ is the same as the associated seed sequence of the path from $\Pi_1(q)$ to $B_{\tau,i}$.

As a result, if one of the lines $L \in R_\tau$ interacts differently with $A_{\tau,i}$ and $B_{\tau,i}$ (e.g. the parity of $|L \cap A_{\tau,i}|$ and $|L \cap B_{\tau,i}|$ are different), for some $i \in [m_\tau]$, then it also distinguishes p, q and we succeed. We also point out that the branching structure of $\Pi(p)$ mimics label propagation in the refinement process: A point in $A_{\tau,i}$, for example may affect the label of p after $2(\tau-1)$ steps by propagating along a sequence of $\tau-1$ lines related to the unique seed sequence of the path from $\Pi_1(p)$ to $A_{\tau,i}$.

The main technical challenge is to design the construction of each level of $\Pi(p)$ and $\Pi(q)$ from the level just built, and to formulate an inductive condition on each level $(\Pi_{\ell-1}(p), \Pi_{\ell-1}(q))$ that allows us to probabilistically grow $\Pi(p), \Pi(q)$ by building $(\Pi_\ell(p), \Pi_\ell(q))$ from $(\Pi_{\ell-1}(p), \Pi_{\ell-1}(q))$ and $R_{\ell-1}$. This condition, when applied on the τth level $(\Pi_\tau(p)$, $\Pi_\tau(q))$, should give R_τ a sufficiently large chance to interact them differently, so that p and q are distinguished.

3.4 A canonical pairwise distinguisher for Steiner-2 designs

We now describe the main technical algorithms, **Cert** and **Test**, for Steiner-2 Designs. As described in last section, we will use both point seeding map and line seeding map. Hence, we define **Cert** $(\mathcal{S}, f, g, p, q)$ and **Test** $(\mathcal{S}, f, g, M, p)$ for Steiner-2 designs as following:

1. The five input parameters of **Cert** include a Steiner 2-design $\mathcal{S} = (\mathcal{P}, \mathcal{L})$, a point-seeding map $f : [t] \to \mathcal{P}$, a line-seeding map $g : [t] \to \mathcal{L}$ for some integer $t \geq 1$, and two distinct points $p, q \in \mathcal{P}$.

2. The five input parameters of **Test** are the same, except that M is a binary string and there is only one point $p \in \mathcal{P}$.

The first algorithm **Cert** $(\mathcal{S}, f, g, p, q)$ iteratively applies two operations Contract and Expand to build two multi-stage branching structures, one for p and one for q, and uses the third operation Interact to certify the pairwise distinction at the final stage. When it succeeds, **Cert** returns a certificate M that is essentially a sketch of the structure built for p. The second algorithm **Test** $(\mathcal{S}, f, g, M, p)$ then applies the same operations Contract, Expand and Interact, trying to build a multi-stage structure for p that matches the description given in M. It outputs 1 if it succeeds; and 0 if it fails.

We will use the following definitions in our algorithms. Let $\mathcal{S} = (\mathcal{P}, \mathcal{L})$ be a Steiner 2-design.

Definition 3.4.1 ((m, W)-PDST)**.** *For positive integers m and W, an* (m, W)-PDST (pairwise disjoint set tuple) *over $\mathcal{P}$ is a tuple $\mathcal{A} = (A_1, \ldots, A_m)$ of subsets of $\mathcal{P}$ that satisfies the following properties:* (i) *the A_i's are pairwise disjoint and nonempty subsets of $\mathcal{P}$; and* (ii) *$W/2 \leq |A_i| \leq W$ for every $i \in [m]$.*

Definition 3.4.2 ((m, W, α)-pair)**.** *For $\alpha \in [0, 1]$, two (m, W)-PDSTs $\mathcal{A}$ and $\mathcal{B}$ form an* (m, W, α)-pair $(\mathcal{A}, \mathcal{B})$ *over $\mathcal{P}$, if $|A_i| = |B_i|$ and $|A_i \cap B_i| \leq \alpha \cdot |A_i|$ for every $i \in [m]$.*

Below, we first define these three operations Contract, Expand and Interact, and state their properties that will be crucially used in algorithms **Cert** and **Test**. We then present **Cert** and **Test**. We prove all technical lemmas and theorems in Section 3.5.

3.4.1 Contraction

Let $\mathcal{A} = (A_1, \ldots, A_m)$ be an (m, W)-PDST over $\mathcal{P}$. Let m', r and W' be three positive integers. Syntactically, we say $\mathcal{C}$ is an (m, m', r, W')-*contraction* map if it satisfies the following two properties:

- For each $j \in [m]$, $\mathcal{C}(j)$ is either nil or a pair (i_j, k_j), where $i_j \in [r]$ and k_j is a positive integer between $W'/2$ and W'; and the number of $j \in [m]$ such that $\mathcal{C}(j) \neq \text{nil}$ is equal to m'.

Now let $\mathbf{L} = (L_1, \ldots, L_r)$ denote a tuple of r not necessarily distinct lines drawn from $\mathcal{L}$. Semantically, we say $\mathcal{C}$ *matches* $(\mathcal{A}, \mathbf{L})$ if for any $j \in [m]$ such that $(i_j, k_j) = \mathcal{C}(j) \neq \text{nil}$, we have $|L_{i_j} \cap A_j| = k_j$.

We use $\mathsf{Contract}(\mathcal{S}, \mathcal{A}, \mathbf{L}, \mathcal{C})$ to denote the following polynomial-time procedure: If $\mathcal{C}$ matches $(\mathcal{A}, \mathbf{L})$, then $\mathcal{A}' = \mathsf{Contract}(\mathcal{S}, \mathcal{A}, \mathbf{L}, \mathcal{C})$ is the tuple consists of m sets $L_{i_j} \cap A_j$, where $(i_j, k_j) = \mathcal{C}(j) \neq \text{nil}$, ordered by j from small to large. Otherwise, we set $\mathsf{Contract}(\mathcal{S}, \mathcal{A}, \mathbf{L}, \mathcal{C}) = \text{nil}$. It is clear that if $\mathcal{C}$ matches $(\mathcal{A}, \mathbf{L})$, then the output $\mathcal{A}'$ of $\mathsf{Contract}$ must be an (m', W')-PDST.

Let $(\mathcal{A}, \mathcal{B})$ be an (m, W, α)-pair and $\mathbf{L} = (L_1, \ldots, L_r)$ be a tuple of r lines. We say an (m, m', r, W')-contraction tuple $\mathcal{C}$ is α'-good with respect to $(\mathcal{A}, \mathcal{B})$ and $\mathbf{L}$, for some $\alpha' \in [0, 1]$, if

- $\mathcal{C}$ matches both $(\mathcal{A}, \mathbf{L})$ and $(\mathcal{B}, \mathbf{L})$; and $(\mathcal{A}', \mathcal{B}')$ is an (m', W', α')-pair, where

$$\mathcal{A}' = \mathsf{Contract}(\mathcal{S}, \mathcal{A}, \mathbf{L}, \mathcal{C}) \quad \text{and} \quad \mathcal{B}' = \mathsf{Contract}(\mathcal{S}, \mathcal{B}, \mathbf{L}, \mathcal{C})$$

  denote the two (m', W')-PDSTs obtained from applying $\mathsf{Contract}$ on $\mathcal{A}$ and $\mathcal{B}$, respectively.

We will prove the following technical lemma in Section 3.5.1:

Lemma 3.4.3 (Contraction)**.** *Let $\mathcal{S} = (\mathcal{P}, \mathcal{L})$ be a Steiner 2-design with parameters (v, n, s, h), where v is bounded below by a sufficiently large constant. Let $(\mathcal{A}, \mathcal{B})$ be an (m, W, α)-pair over $\mathcal{P}$ with*

$$\beta = \frac{s}{v} \cdot W \geq 1/8$$

Let ϵ, γ and r be the following three parameters:

$$\epsilon = \frac{1}{\lceil \log v \rceil}, \quad \gamma = \frac{\epsilon^2}{264} \quad \textit{and} \quad r = \frac{4\lceil \log v \rceil}{\gamma} = 1056 \cdot \lceil \log v \rceil^3$$

If r lines $\mathbf{L} = (L_1, \ldots, L_r)$ are sampled uniformly at random, then with probability at least $1 - 1/v^3$, one can construct in polynomial time an (m, m', r, W')-contraction map $\mathcal{C}$ such that

$$m' = \lceil m / \lceil \log v \rceil \rceil \quad \textit{and} \quad W' \leq \beta/\gamma$$

*and one of the following two conditions holds: either **a)** $\mathcal{C}$ is α'-good with respect to $(\mathcal{A}, \mathcal{B})$ and $\mathbf{L}$, where $\alpha' = (1+\epsilon)\alpha$; or **b)** $\mathcal{C}$ matches $(\mathcal{A}, \mathbf{L})$ but does not match $(\mathcal{B}, \mathbf{L})$.*

3.4.2 Expansion

Assume $s \geq 3$. From now on we let $s' = s - 2 \geq 1$. In Expand, we will use the following definition:

Definition 3.4.4 (Cones)**.** *Let $T \subset \mathcal{P}$ and $p \in \mathcal{P}$. We use $\textsc{cone}(T, p) \subset \mathcal{P}$ to denote the following set, which will be referred to as the* cone *defined by T and p: When $p \in T$, $\textsc{cone}(T, p) = \emptyset$; When $p \notin T$,*

$$\textsc{cone}(T, p) = \Big\{ q \notin T \cup \{p\} : \textit{the line } L \in \mathcal{L} \textit{ uniquely determined by } p \textit{ and } q \textit{ satisfies } |L \cap T| = 1 \Big\}.$$

It is clear from the definition that $\textsc{cone}(T, p) \cap T = \emptyset$ and $p \notin \textsc{cone}(T, p)$.

Definition 3.4.5. *Let $T \subset \mathcal{P}$ and $p \in \mathcal{P}$. We say $q \in T$ is a* good *point with respect to (T, p) if $p \notin T$ and the line $L \in \mathcal{L}$ uniquely determined by p and q intersects with T only at q. Otherwise, we say $q \in T$ is a* bad *point with respect to (T, p). Note that when $p \in T$, every point in T is bad with respect to (T, p). We use $\textsc{gp}(T, p)$ and $\textsc{bp}(T, p)$ to denote the set of all good and bad points in T, respectively, with respect to (T, p).*

From these definitions, it is easy to prove the following two lemmas:

Lemma 3.4.6. *Let $T \subset \mathcal{P}$ and $p \in \mathcal{P}$, then we have*

$$\textsc{cone}(T, p) = \textsc{cone}(\textsc{gp}(T, p), p) \quad \textit{and} \quad |\textsc{cone}(T, p)| = s' \cdot |\textsc{gp}(T, p)|$$

Lemma 3.4.7. *Let $T \subset \mathcal{P}, p \in \mathcal{P}$ and $H \subseteq \text{GP}(T, p)$ be a subset of good points with respect to (T, p), then*

$$|\text{CONE}(H, p)| = s' \cdot |H|$$

Now we define Expand. Let $\mathcal{A} = (A_1, \ldots, A_m)$ denote an (m, W)-PDST over $\mathcal{P}$. Let m', r, and W' be three positive integers. Syntactically, we call $\mathcal{E}$ an (m, m', r, W')-*expansion map* if

- For any $i \in [r]$ and $j \in [m]$, $\mathcal{E}(i, j)$ is either nil or a positive integer $k_{i,j}$ between $W'/2$ and W'; and the number of (i, j) such that $k_{i,j} = \mathcal{E}(i, j) \neq \text{nil}$ is exactly m'.

Let $A = \cup_{j \in [m]} A_j$, and let $\mathbf{p} = (p_1, \ldots, p_r)$ denote a tuple of r not necessarily distinct points drawn from $\mathcal{P}$. For each pair $(i, j) \in [r] \times [m]$, we let $G_{i,j} \subseteq A_j$ denote the set of good points in A_j with respect to (A, p_i): $G_{i,j} = A_j \cap \text{GP}(A, p_i)$. Then semantically we say $\mathcal{E}$ matches $(\mathcal{A}, \mathbf{p})$ if for every $(i, j) \in [r] \times [m]$ such that $k_{i,j} = \mathcal{E}(i, j) \neq \text{nil}$, we have

$$|A_{i,j}| = k_{i,j}, \quad \text{where } A_{i,j} = \text{CONE}(G_{i,j}, p_i) - \cup_{\ell < i} \text{CONE}(A, p_\ell)$$

Now we can define the operation Expand. If $\mathcal{E}$ matches $(\mathcal{A}, \mathbf{p})$, then we set

$$\mathcal{A}' = \mathsf{Expand}(\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathcal{E})$$

to be the tuple of $A_{i,j}$'s, where $\mathcal{E}(i, j) \neq \text{nil}$, ordered by (i, j) under the lexicographical order. It is clear that when $\mathcal{E}$ matches $(\mathcal{A}, \mathbf{p})$, $\mathcal{A}'$ is an (m', W')-PDST. If they do not match then $\mathsf{Expand}(\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathcal{E}) = \text{nil}$.

Finally, let $(\mathcal{A}, \mathcal{B})$ denote an (m, W, α)-pair. We say an (m, m', r, W')-expansion map $\mathcal{E}$ is α'-good with respect to $(\mathcal{A}, \mathcal{B})$ and $\mathbf{p}$, for some $\alpha' \in [0, 1]$, if the following two conditions hold:

- $\mathcal{E}$ matches both $(\mathcal{A}, \mathbf{p})$ and $(\mathcal{B}, \mathbf{p})$; and $(\mathcal{A}', \mathcal{B}')$ is an (m', W', α')-pair, where

  $$\mathcal{A}' = \mathsf{Expand}(\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathcal{E}) \quad \text{and} \quad \mathcal{B}' = \mathsf{Expand}(\mathcal{S}, \mathcal{B}, \mathbf{p}, \mathcal{E})$$

  denote the two (m', W')-PDSTs obtained from applying Expand on $\mathcal{A}$ and $\mathcal{B}$, respectively.

We will prove the following technical lemma in Section 3.5.2.

Lemma 3.4.8 (Expansion)**.** *Let $\mathcal{S} = (\mathcal{P}, \mathcal{L})$ be a Steiner 2-design with parameters (v, n, s, h), where $s \geq 3$ and v is bounded below by a sufficiently large constant. Let ϵ and r denote*

$$r = 2^{10} \cdot \lceil \log v \rceil \quad \textit{and} \quad \epsilon = \frac{1}{2^{17} \cdot \lceil \log v \rceil^2}$$

Let $(\mathcal{A}, \mathcal{B})$ be an (m, W, α)-pair over $\mathcal{P}$ with

$$2 \cdot \frac{s}{v} \cdot m \cdot W \leq \epsilon \quad \textit{and} \quad \alpha \leq 1/2$$

If r points $\mathbf{p} = (p_1, \ldots, p_r)$ are sampled uniformly at random from $\mathcal{P}$, then with probability at least $1 - 1/v^{64}$, one can construct in polynomial time an (m, m', r, W')-expansion map $\mathcal{E}$ such that

$$m' \geq m \cdot 32 \lceil \log v \rceil$$

*and one of the following two conditions holds: either **a)** $\mathcal{E}$ is α'-good with respect to $(\mathcal{A}, \mathcal{B})$ and $\mathbf{p}$ with*

$$\alpha' = \alpha + \frac{1}{2 \log v}$$

*or **b)** $\mathcal{E}$ matches $(\mathcal{A}, \mathbf{p})$ but does not match $(\mathcal{B}, \mathbf{p})$.*

3.4.3 Interaction

Let $\mathcal{A} = (A_1, \ldots, A_m)$ denote an (m, W)-PDST over $\mathcal{P}$, and let $\mathbf{L} = (L_1, \ldots, L_r)$ denote a tuple of r lines drawn from $\mathcal{L}$. Given any $i \in [r]$ and $j \in [m]$, $\mathsf{Interact}(\mathcal{S}, \mathcal{A}, \mathbf{L}, (i, j))$ just returns the parity of $|L_i \cap A_j|$ (0 if it is even and 1 if it is odd). We will prove the following technical lemma in Section 3.5.3:

Lemma 3.4.9. *Let $\mathcal{S} = (\mathcal{P}, \mathcal{L})$ be a Steiner 2-design with parameters (v, n, s, h), where v is bounded below by a sufficiently large constant. Let $(\mathcal{A}, \mathcal{B})$ denote an (m, W, α)-pair with*

$$\alpha \leq \frac{1}{2}, \quad \frac{s}{v} \cdot W \leq \frac{1}{8} \quad \textit{and} \quad \frac{s}{v} \cdot m \cdot W \geq \frac{1}{2^{18} \cdot \lceil \log v \rceil^2}$$

Let $r = 2^{23} \cdot \lceil \log v \rceil^3$. Then if r lines $\mathbf{L} = (L_1, \ldots, L_r)$ are sampled uniformly at random from $\mathcal{L}$, then with probability at least $1 - 1/v^4$, there exists a pair $i \in [r]$ and $j \in [m]$ such that

$$\mathsf{Interact}(\mathcal{S}, \mathcal{A}, \mathbf{L}, (i, j)) \neq \mathsf{Interact}(\mathcal{S}, \mathcal{B}, \mathbf{L}, (i, j)) \tag{3.3}$$

3.4.4 Algorithms Cert and Test

The goal of algorithm **Cert** $(\mathcal{S}, f, g, p, q)$ is to produce a certificate M that will assist algorithm **Test** to distinguish $p, q \in \mathcal{P}$. If successful, **Cert** $(\mathcal{S}, f, g, p, q)$ will produce a certificate of the following form.

Definition 3.4.10 (Certificates)**.** *A* certificate of a pairwise distinguisher M *is a finite tuple in which each component is either a* contraction map $\mathcal{C}$*, an* expansion map $\mathcal{E}$*, or a pair of positive integers* (i, j)*.*

In Figure 3.1 we present algorithm **Test** $(\mathcal{S}, f, g, M, p)$ whose input parameters are: a Steiner-2 design $\mathcal{S} = (\mathcal{P}, \mathcal{L})$, two maps $f : [t] \to \mathcal{P}$ and $g : [t] \to \mathcal{L}$, a certificate of a pairwise distinguisher M, and a point $p \in \mathcal{P}$. It also uses the following parameters. Assume $\mathcal{S}$ has parameters (v, n, s, h), where $s \geq 3$ and v is bounded below by a sufficiently large constant. Let r_1, r_2 and r_3 denote the following positive integers:

$$r_1 = 1056 \cdot \lceil \log v \rceil^3, \qquad r_2 = 2^{10} \cdot \lceil \log v \rceil \qquad \text{and} \qquad r_3 = 2^{23} \cdot \lceil \log v \rceil^3 \tag{3.4}$$

In addition to r_1, r_2 and r_3 above, algorithm **Cert**, given in Figure 3.2, uses an additional parameter:

$$\epsilon = \frac{1}{2^{17} \cdot \lceil \log v \rceil^2}$$

To simplify our presentation of **Cert**, we use a '**case-when** construct' which executes the block of statements that follows the *first* 'case-when' Boolean expression that is true. In other words, in each iteration of **Cert** as given in Figure 3.2, we perform either *Expansion*, or *Contraction*, or *Interaction*, with *Expansion* preferred over *Contraction* when both of their Boolean expressions are true.

We prove following theorem for **Cert**. By Corollary 2.2.6, Theorem 3.2.2 follows.

Theorem 3.4.11 (Polylogarithmic Number of Seeds Suffice)**.** *There exist three positive constants* C_1, C_2 *and* C_3 *such that for any Steiner* 2*-design* $\mathcal{S} = (\mathcal{P}, \mathcal{L})$ *with parameters* (v, n, s, h) *satisfying*

$$v \geq C_1, \quad s \geq 3, \quad \textit{and} \quad n \geq (C_2 \cdot \log^3 v)\, sh$$

Algorithm: Test $(\mathcal{S}, f, g, M, p)$

1. Let K denote the number of components in M
2. Set $\mathcal{A}_0 = (\{p\})$, i.e., a $(1,1)$-PDST over $\mathcal{P}$; and set $a = b = 0$
3. For k from 1 to K do
4. If the kth component of M is an (m, m', r, W')-contraction map $\mathcal{C}$ then
5. If $b + r_1 > t$ then return 0 (% `running out random lines`)
6. else if $\mathcal{C}$ does not match $\big(\mathcal{A}_{k-1}, (g(b+1), \ldots, g(b+r_1))\big)$ then return 0
7. else: set $\mathcal{A}_k = \mathsf{Contract}\big(\mathcal{S}, \mathcal{A}_{k-1}, (g(b+1), \ldots, g(b+r_1)), \mathcal{C}\big)$ and set $b = b + r_1$
8. else if the kth component of M is an (m, m', r, W')-expansion map $\mathcal{E}$ then
9. If $a + r_2 > t$ then return 0 (% `running out random points`)
10. else if $\mathcal{E}$ does not match $\big(\mathcal{A}_{k-1}, (f(a+1), \ldots, f(a+r_2))\big)$ then return 0
11. else: set $\mathcal{A}_k = \mathsf{Expand}\big(\mathcal{S}, \mathcal{A}_{k-1}, (f(a+1), \ldots, f(a+r_2)), \mathcal{E}\big)$ and set $a = a + r_2$
12. else: the kth component of M is a pair (i, j) of integers
13. If $b + r_3 > t$ then return 0 (% `running out random lines`)
14. else return $\mathsf{Interact}\big(\mathcal{S}, \mathcal{A}_{k-1}, (g(b+1), \ldots, g(b+r_3)), (i, j)\big)$
15. End for
16. return 1

Figure 3.1: Description of the algorithm **Test** for Steiner 2-designs

there exists a pair of maps, $f^* : [t] \to \mathcal{P}$ *and* $g^* : [t] \to \mathcal{L}$*, with* $t = \lceil C_3 \cdot \log^4 v \rceil$*, that guarantees*

$$\mathbf{Cert}\,(\mathcal{S}, f^*, g^*, p, q) \neq \text{nil}, \quad \textit{for any two distinct points } p, q \in \mathcal{P}.$$

Before moving to the analysis of the three operations and finally the proof of Theorem 3.4.11, we first show that (**Cert**, **Test**) forms a canonical pairwise distinguisher for Steiner-2 designs.

Proof of Property 4.2.29. First we can prove that both algorithms **Cert** and **Test** are invariant under isomorphisms by a routine induction following the loops of **Cert** and **Test**, and also using the fact that all three operations Contract, Expand and Interact are themselves invariant under isomorphism.

Second, to see that (**Cert**, **Test**) satisfies the condition of pairwise distinctness, we consider any two points $p, q \in \mathcal{P}$. If $M = \mathbf{Cert}\,(\mathcal{S}, f, g, p, q) \neq \text{nil}$, then M is a certificate, and its last element is either a contraction map $\mathcal{C}$, or an expansion map $\mathcal{E}$, or a pair of integers (i, j). Per algorithm **Cert**, its output is not nil only when the last element of M produces a mismatch between the $\mathcal{A}$ structure and $\mathcal{B}$ structure, constructed from p and q, respectively. As a result, when algorithm **Test** traces the branching structure to this element, it will return different Boolean values for p and q. □

3.5 Analysis

In the analysis of this section, we always assume that $\mathcal{S} = (\mathcal{P}, \mathcal{L})$ is a Steiner 2-design with parameters (v, n, s, h), where $s \geq 3$ and v is bounded below by a sufficiently large constant.

The following property of Steiner-2 designs will be very useful to our analysis.

Proposition 3.5.1 (Points and Lines)**.** *For any subset of points* $P \subseteq \mathcal{P}$*, we have*

$$\sum_{L \in \mathcal{L}} |L \cap P| = h|P| \quad \textit{and} \quad \sum_{L \in \mathcal{L}} |L \cap P| \cdot \big(|L \cap P| - 1\big) = |P| \cdot \big(|P| - 1\big)$$

Here the first equation follows from the definition that every point $p \in \mathcal{P}$ belongs to exactly h lines, and the second equation is true because for every two distinct points $p, q \in \mathcal{P}$, there exists a unique line $L \in \mathcal{L}$ such that $p, q \in L$.

Algorithm: Cert $(\mathcal{S}, f, g, p, q)$:

1. [*Initialization*] Set $\mathcal{A}_0 = (\{p\})$ and $\mathcal{B}_0 = (\{q\})$; set $(m_0, W_0, \alpha_0) = (1, 1, 0)$
2. Set M to be the empty tuple; and set $a = b = 0$
3. For k from 1 to $\lfloor \log v/2 \rfloor$ do
4. [*Comment*] $(\mathcal{A}_{k-1}, \mathcal{B}_{k-1})$ is an $(m_{k-1}, W_{k-1}, \alpha_{k-1})$-pair over $\mathcal{P}$ (by induction)
5. **Case** [*Expansion*]: **when** $2(s/v) \cdot m_{k-1} \cdot W_{k-1} \leq \epsilon$
6. If $a + r_2 > t$ then return nil (% `running out random points`)
7. set $\mathbf{p} = \big(f(a+1), \ldots, f(a+r_2)\big)$ and $a = a + r_2$
8. find an (m_{k-1}, m_k, r_2, W_k)-expansion map $\mathcal{C}$ satisfying the conditions of Lemma 3.4.8
9. If Lemma 3.4.8 fails then return nil
10. else if $\mathcal{C}$ satisfies ***b)*** of Lemma 3.4.8, then add $\mathcal{C}$ to the end of M and return M
11. else add $\mathcal{C}$ to the end of M; $\mathcal{A}_k = \mathsf{Expand}\big(\mathcal{S}, \mathcal{A}_{k-1}, \mathbf{L}, \mathcal{C}\big)$; $\mathcal{B}_k = \mathsf{Expand}\big(\mathcal{S}, \mathcal{B}_{k-1}, \mathbf{L}, \mathcal{C}\big)$
12. **Case** [*Contraction*]: **when** $(s/v) \cdot W_{k-1} \geq 1/8$
13. If $b + r_1 > t$ then return nil (% `running out random lines`)
14. set $\mathbf{L} = \big(g(b+1), \ldots, g(b+r_1)\big)$ and $b = b + r_1$
15. find an (m_{k-1}, m_k, r_1, W_k)-contraction map $\mathcal{E}$ satisfying the conditions of Lemma 3.4.3
16. If Lemma 3.4.3 fails then return nil
17. else if $\mathcal{E}$ satisfies ***b)*** of Lemma 3.4.3, then add $\mathcal{E}$ to the end of M and return M
18. else add $\mathcal{E}$ to the end of M; $\mathcal{A}_k = \mathsf{Contract}\big(\mathcal{S}, \mathcal{A}_{k-1}, \mathbf{L}, \mathcal{E}\big)$; $\mathcal{B}_k = \mathsf{Contract}\big(\mathcal{S}, \mathcal{B}_{k-1}, \mathbf{L}, \mathcal{E}\big)$
19. **Case** [*Interaction*]: **when** $(s/v) \cdot W_{k-1} < 1/8$ and $2(s/v) \cdot m_{k-1} \cdot W_{k-1} > \epsilon$
20. If $b + r_3 > t$ then return nil (% `running out random lines`)
21. set $\mathbf{L} = \big(g(b+1), \ldots, g(b+r_3)\big)$ and $b = b + r_3$
22. find a pair $(i, j) \in [r_3] \times [m_{k-1}]$ of integers that satisfies the conditions of Lemma 3.4.9
23. If Lemma 3.4.9 fails then return nil
24. else add (i, j) to the end of M and return M
25. End for
26. [*Comment*] This line should never be reached

Figure 3.2: Description of the algorithm **Cert** for Steiner 2-designs

3.5.1 Contraction: Proof of Lemma 3.4.3

We first establish the following two lemmas, which will be used in the proof of Lemma 3.4.3.

Lemma 3.5.2. *Let P and Q be two subsets of $\mathcal{P}$, where P is nonempty but Q could be empty. Let*

$$\beta = \frac{s}{v} \cdot |P| > 0 \quad \textit{and} \quad \alpha = \frac{|Q|}{|P|} \geq 0$$

Let $\epsilon \in (0, 1]$ be a parameter. If a line L is sampled uniformly at random from $\mathcal{L}$, then

$$\Pr\Big[\, |L \cap P| \geq 1 \textit{ and } |L \cap Q| \leq (1+\epsilon)\alpha \cdot |L \cap P| \,\Big] \geq \frac{\epsilon^2 \beta}{4(2+\beta)}. \tag{3.5}$$

Proof. Let $\mathcal{L}^*$ denote the set of lines $L \in \mathcal{L}$ such that $|L \cap Q| > (1+\epsilon)\alpha \cdot |L \cap P|$. By Proposition 4.12,

$$h|Q| = \sum_{L \in \mathcal{L}} |L \cap Q| \geq \sum_{L \in \mathcal{L}^*} |L \cap Q| > (1+\epsilon) \cdot \frac{|Q|}{|P|} \cdot \sum_{L \in \mathcal{L}^*} |L \cap P| \implies \sum_{L \in \mathcal{L}^*} |L \cap P| < \frac{h|P|}{1+\epsilon}$$

Since $\sum_{L \in \mathcal{L}} |L \cap P| = h|P|$ by Proposition 4.12, we have

$$\sum_{L \notin \mathcal{L}^*} |L \cap P| > h|P| \cdot \frac{\epsilon}{1+\epsilon} \geq \frac{\epsilon h|P|}{2}$$

where the last inequality uses $\epsilon \leq 1$.

Now let $\mathcal{L}'$ denote the set of lines $L \in \mathcal{L}$ such that $|L \cap P| \geq 1$ and $L \notin \mathcal{L}^*$. Then we have

$$\Pr\Big[\, |L \cap P| \geq 1 \text{ and } |L \cap Q| \leq (1+\epsilon)\alpha \cdot |L \cap P| \,\Big] = \frac{|\mathcal{L}'|}{n}. \tag{3.6}$$

To prove the lemma, we give a lower bound for $|\mathcal{L}'|$. First we note that

$$\sum_{L \in \mathcal{L}'} |L \cap P| = \sum_{L \notin \mathcal{L}^*} |L \cap P| > \frac{\epsilon h|P|}{2} \tag{3.7}$$

Then by Cauchy-Schwarz, we have

$$\left(\frac{\epsilon h|P|}{2}\right)^2 < \left(\sum_{L \in \mathcal{L}'} |L \cap P|\right)^2 \leq |\mathcal{L}'| \cdot \left(\sum_{L \in \mathcal{L}'} |L \cap P|^2\right) \tag{3.8}$$

Using the second equation of Proposition 4.12, we get

$$\sum_{L \in \mathcal{L}'} |L \cap P|^2 \leq \sum_{L \in \mathcal{L}} |L \cap P| \cdot \big(|L \cap P| - 1\big) + \sum_{L \in \mathcal{L}} |L \cap P| = |P|\big(|P| - 1\big) + h|P| < |P|^2 + h|P|$$

Plugging this in (3.8), we get the following lower bound for $|\mathcal{L}'|$:

$$|\mathcal{L}'| > \frac{1}{|P|^2 + h|P|} \cdot \left(\frac{\epsilon h |P|}{2}\right)^2 = \frac{\epsilon^2}{4} \cdot \frac{h^2|P|}{|P| + h}$$

Moreover, because $(s/v) \cdot |P| = \beta$ and (3.2), we have

$$|P| = \beta \cdot \frac{v}{s} \geq \beta \cdot \frac{h}{2} \implies |\mathcal{L}'| > \frac{\epsilon^2}{4} \cdot \frac{h^2|P|}{|P| \cdot (1 + 2/\beta)} = \frac{\epsilon^2 \beta}{4(2+\beta)} \cdot h^2 \geq \frac{\epsilon^2 \beta}{4(2+\beta)} \cdot n$$

The lemma then follows from (3.6). □

Lemma 3.5.3 (Concentration of Line Intersection)**.** *Let $\gamma \in (0,1)$ be a parameter. Let P be a nonempty subset of $\mathcal{P}$ with $\beta = (s/v) \cdot |P| > 0$. If a line L is sampled uniformly at random, then we have*

$$\Pr\Big[\, |L \cap P| \leq \beta/\gamma \,\Big] \geq 1 - \gamma \tag{3.9}$$

Proof. For each $k \in [0:s]$, let N_k denote the number of lines $L \in \mathcal{L}$ such that $|L \cap P| = k$. Then

$$\sum_k N_k = n \quad \text{and} \quad \sum_k N_k \cdot k = h|P| = \frac{n \cdot s}{v} \cdot |P| = \beta n$$

by Proposition 4.12 and (3.1). We then have

$$\beta n \geq \sum_{k > \beta/\gamma} N_k \cdot k > \frac{\beta}{\gamma} \cdot \sum_{k > \beta/\gamma} N_k \implies \sum_{k > \beta/\gamma} N_k < \gamma n \implies \sum_{k \leq \beta/\gamma} N_k > (1-\gamma) n$$

Therefore, the probabilty that $|L \cap P| \leq \beta/\gamma$ is at least $1 - \gamma$. □

Combining these two lemmas, we get the following useful corollary:

Corollary 3.5.4. *Let P and Q be two subsets of $\mathcal{P}$ with*

$$\beta = \frac{s}{v} \cdot |P| \geq \frac{1}{16} \quad \textit{and} \quad \alpha = \frac{|Q|}{|P|} \geq 0$$

Let $\epsilon, \gamma \in (0,1)$ be two parameters such that $\gamma = \epsilon^2/264$. Then we have

$$\Pr\Big[\, 1 \leq |L \cap P| < \beta/\gamma \textit{ and } |L \cap Q| \leq (1+\epsilon)\alpha \cdot |L \cap P| \,\Big] \geq \gamma$$

Proof. By Lemma 3.5.3, we have (3.9). By Lemma 3.5.2, we have

$$\Pr\Big[\, |L \cap P| \geq 1 \text{ and } |L \cap Q| \leq (1+\epsilon)\alpha \cdot |L \cap P| \,\Big] \geq \frac{\epsilon^2 \beta}{4(2+\beta)} \geq \frac{\epsilon^2}{132} = 2\gamma$$

where the second inequality uses $\beta \geq 1/16$. The lemma then follows from the union bound.

□

We are now ready to prove Lemma 3.4.3.

Proof of Lemma 3.4.3. Let $H_j = A_j \cap B_j$ for each $j \in [m]$. Then $|H_j| \le \alpha \cdot |A_j|$. Because

$$\frac{s}{v} \cdot |A_j| \ge \frac{s}{v} \cdot \frac{W}{2} \ge \frac{1}{16}$$

by Corollary 3.5.4 we have for any $j \in [m]$ and $i \in [r]$:

$$\Pr\left[\, 0 < |L_i \cap A_j| < \beta/\gamma \text{ and } |L_i \cap H_j| \le (1+\epsilon) \cdot \alpha \cdot |L_i \cap A_j| \,\right] \ge \gamma.$$

As a result, we have for each $j \in [m]$:

$$\begin{aligned}\Pr\left[\, \exists\, i \in [r] \text{ such that } 0 < |L_i \cap A_j| < \beta/\gamma \text{ and } |L_i \cap H_j| \le (1+\epsilon) \cdot \alpha \cdot |L_i \cap A_j| \,\right] \\ \ge 1 - (1-\gamma)^r \ge 1 - \exp(-\gamma r) \ge 1 - 1/v^4.\end{aligned}$$

Since $m \le v$, by the union bound we have

$$\begin{aligned}\Pr[\forall\, j \in [m],\, \exists\, i \in [r] \text{ such that } 0 < |L_i \cap A_j| < \beta/\gamma \\ \text{and } |L_i \cap H_j| \le (1+\epsilon) \cdot \alpha \cdot |L_i \cap A_j| \,] \ge 1 - 1/v^3.\end{aligned}$$

Assume that the event above happens: For every $j \in [m]$, there is an $i_j \in [r]$ such that

$$0 < |L_{i_j} \cap A_j| < \beta/\gamma \quad \text{and} \quad |L_{i_j} \cap H_j| \le (1+\epsilon) \cdot \alpha \cdot |L_{i_j} \cap A_j|$$

We then construct $\mathcal{C}$ as follows: Divide $[1:s]$ into $\lceil \log(s+2) \rceil - 1 \le \lceil \log s \rceil \le \lceil \log v \rceil$ many intervals:

$$[1:2],\ [3:6],\ [7:14],\ \ldots,\ [2^i - 1 : 2^{i+1} - 2],\ \ldots$$

and let I denote the interval that contains the most $j \in [m]$ such that $|L_{i_j} \cap A_j| \in I$. It is clear that the number of $j \in [m]$ such that $|L_{i_j} \cap A_j| \in I$ is at least $m' = \lceil m / \lceil \log v \rceil \rceil$. Then we set

$$\mathcal{C}(j) = \big(i_j, |L_{i_j} \cap A_j|\big)$$

if $|L_{i_j} \cap A_j| \in I$ and j is one of the m' smallest such $j \in [m]$; and we set $\mathcal{C}(j) = \text{nil}$ otherwise. It is easy to check that $\mathcal{C}$ is an (m, m', r, W')-contraction map for some appropriate positive integer $W' \le \beta/\gamma$. It satisfies either condition ***a)*** or condition ***b)*** of Lemma 3.4.3, based on whether $\mathcal{C}$ matches $(\mathcal{B}, \mathbf{L})$. □

3.5.2 Expansion: Proof of Lemma 3.4.8

We start with the following two technical lemmas:

Lemma 3.5.5. *Let $\epsilon, \gamma \in (0,1)$ be two parameters and $T \subset \mathcal{P}$ be a set of points satisfying $(s/v) \cdot |T| < \epsilon$. If a point p is sampled uniformly at random, then we have*

$$\Pr\Big[\, p \notin T \text{ and } \big|\,\mathrm{BP}\,(T,p)\,\big| \le \frac{\epsilon}{\gamma} \cdot |T| \,\Big] \ge 1 - \gamma - \epsilon \tag{3.10}$$

Proof. For each $k \in [0:s]$, let N_k be the number of lines L such that $|L \cap T| = k$. By Proposition 4.12,

$$\sum_k N_k \cdot k(k-1) = |T| \cdot \big(|T| - 1\big) < |T|^2$$

For each point $p \in \mathcal{P}$ and $k \in [0:s]$, we let $N_{k,p}$ denote the number of lines $L \in \mathcal{L}$ such that $p \in L$ and $|L \cap T| = k$. If p is sampled uniformly at random, then we have

$$\mathbf{E}_p\left[\sum_k N_{k,p} \cdot k(k-1)\right] = \frac{s}{v} \cdot \sum_k N_k \cdot k(k-1) < \epsilon |T|$$

By Markov's inequality, with probability at least $1 - \gamma$, we have

$$\sum_k N_{k,p} \cdot k(k-1) \le \frac{\epsilon}{\gamma} \cdot |T| \tag{3.11}$$

It is also easy to see that $p \notin T$ with probability $1 - |T|/v > 1 - \epsilon/s > 1 - \epsilon$. Then, by union bound

$$\Pr\Big[\, p \notin T \text{ and (3.11) holds} \,\Big] \ge 1 - \epsilon - \gamma$$

Now it suffices to show $p \notin T$ and (3.11) together imply (3.10). To see this, from $p \notin T$ we have

$$\textstyle\sum_k N_{k,p} \cdot k = |T|$$

Combining it with (3.11), we have

$$\frac{\epsilon}{\gamma} \cdot |T| \ge \sum_k N_{k,p} \cdot k(k-1) \ge \sum_{k \ge 2} N_{k,p} \cdot k \quad \Longrightarrow \quad N_{1,p} \ge \left(1 - \frac{\epsilon}{\gamma}\right) \cdot |T|$$

The lemma follows because $N_{1,p}$ is exactly the number of good points in T with respect to (T,p). ∎

Recall that $s' = s - 2 \geq 1$. Next we prove the following lemma:

Lemma 3.5.6. *Let $\epsilon \in (0,1)$ be a parameter. Let $A \subset \mathcal{P}$ be a set of points with $(s/v) \cdot |A| \leq \epsilon$, and let $F \subseteq \overline{A}$ be a set of points. If a point p is sampled uniformly at random, then we have*

$$\Pr\Big[\, \big|\, \textsc{cone}\,(A,p) \cap F \,\big| \leq 4\epsilon|F| \,\Big] \geq 3/4 \tag{3.12}$$

Proof. For each point $q \in F$ (and thus $q \notin A$), we let X_q denote the following indicator random variable: $X_q = 1$ when $q \in \textsc{cone}\,(A,p)$; and $X_q = 0$ otherwise. Note that $X_q = 1$ iff $p \in \textsc{cone}\,(A,q)$. Thus,

$$\Pr\big[X_q = 1\big] = \frac{|\,\textsc{cone}\,(A,q)\,|}{v} \leq \frac{s'}{v} \cdot |A| < \frac{s}{v} \cdot |A| \leq \epsilon$$

This implies that the expectation of $\sum_{q \in F} X_q$ is at most $\epsilon|F|$. As a result, we have

$$\Pr\Big[\, \textstyle\sum_{q \in F} X_q \leq 4\epsilon|F| \,\Big] \geq 3/4$$

by Markov's inequality. The lemma then follows directly. □

Combining Lemma 3.5.5 and Lemma 3.5.6, we obtain the following corollary:

Corollary 3.5.7. *Let $\epsilon, \gamma \in (0,1)$ be two parameters, and let $(\mathcal{A}, \mathcal{B})$ be an (m, W, α)-pair over $\mathcal{P}$ with $2(s/v) \cdot mW \leq \epsilon$. Let $A = \cup_i A_i$, $B = \cup_i B_i$, $T = A \cup B$ and let $F \subseteq \overline{A}$ be a set of points. If a point p is sampled uniformly at random from $\mathcal{P}$, then we have*

$$\Pr\big[\,\text{event (3.10) and event (3.12)}\,\big] \geq (3/4) - \epsilon - \gamma$$

Proof. Since $(\mathcal{A}, \mathcal{B})$ is an (m, W, α)-pair over $\mathcal{P}$, we have

$$\frac{s}{v} \cdot |A| \leq \frac{s}{v} \cdot |T| \leq \frac{s}{v} \cdot 2mW \leq \epsilon.$$

The lemma then follows from Lemma 3.5.5 and Lemma 3.5.6 using union bound. □

We can now use Corollary 3.5.7 to prove Lemma 3.4.8.

Proof of Lemma 3.4.8. Let r, ϵ, γ and λ denote the following four parameters:

$$r = 2^{10} \cdot \lceil \log v \rceil, \quad \epsilon = \frac{1}{2^{17} \cdot \lceil \log v \rceil^2}, \quad \gamma = \frac{1}{2^{12} \cdot \lceil \log v \rceil} \quad \text{and} \quad \lambda = 4\epsilon r = \frac{1}{2^5 \cdot \lceil \log v \rceil}$$

Let $A = \cup_j A_j$, $B = \cup_j B_j$ and $T = A \cup B$. For each $(i, j) \in [r] \times [m]$, we use $G_{i,j} \subseteq A_j$ to denote the set of good points in A_j, with respect to (A, p_i): $G_{i,j} = \textsc{gp}(A, p_i) \cap A_j$; and also use $H_{i,j} \subseteq B_j$ to denote the set of good points in B_j, with respect to (B, p_i): $H_{i,j} = \textsc{gp}(B, p_i) \cap B_j$. Then we have

$$\textsc{cone}(A, p_i) = \bigcup_{j \in [m]} \textsc{cone}(G_{i,j}, p_i) \quad \text{and} \quad \textsc{cone}(B, p_i) = \bigcup_{j \in [m]} \textsc{cone}(H_{i,j}, p_i)$$

We now sample $p_1, \ldots, p_r$ one by one. For each $k \in [r]$ and $j \in [m]$, we set $A_{k,j}$ and $B_{k,j}$ as follows:

- Let F_k and F_k^* denote the following two sets of points:

$$F_k = \bigcup_{i<k} \textsc{cone}(A, p_i) \quad \text{and} \quad F_k^* = \bigcup_{i<k} \textsc{cone}(B, p_i)$$

  It is clear that $F_k \cap A = \emptyset$, $F_k^* \cap B = \emptyset$ and $|F_k|, |F_k^*| \leq rs' \cdot |A| = rs' \cdot |B|$. Then for each $j \in [m]$,

$$A_{k,j} = \textsc{cone}(G_{k,j}, p_k) - F_k \quad \text{and} \quad B_{k,j} = \textsc{cone}(H_{k,j}, p_k) - F_k^*$$

By Corollary 3.5.7, we know that, with probability $\geq 3/4 - \epsilon - \gamma > 1/2$, p_k satisfies $p_k \notin T$;

$$\left| \textsc{bp}(T, p_k) \right| \leq \frac{\epsilon}{\gamma} \cdot |T| \quad \text{and} \quad \left| \textsc{cone}(A, p_k) \cap F_k \right| \leq 4\epsilon |F_k| \tag{3.13}$$

From now on, we say p_k is *good* if all three conditions above hold.

Next we show that if p_k is good, then the number of $j \in [m]$ such that $A_{k,j}$ and $B_{k,j}$ satisfy

$$\frac{s'}{2} \cdot |A_j| \leq |A_{k,j}| \leq s' \cdot |A_j| \quad \text{and} \quad |A_{k,j} \cap B_{k,j}| \leq \alpha' \cdot |A_{k,j}|, \quad \text{where } \alpha' = \alpha + \frac{1}{2 \log v} \tag{3.14}$$

is at least $m/4$. To this end, for each $j \in [m]$ we let $G'_{k,j}$ denote $\textsc{gp}(T, p_k) \cap A_j$, the set of good points in A_j, *with respect to* (T, p_k). It is easy to see that a good point in A_j

with respect to (T, p_k) must be good with respect to (A, p_k) as well. This implies that $G'_{k,j} \subseteq G_{k,j}$, $\textsc{cone}\,(G'_{k,j}, p_k) \subseteq \textsc{cone}\,(G_{k,j}, p_k)$ and

$$\textsc{cone}\,(G'_{k,j}, p_k) - F_k \subseteq \textsc{cone}\,(G_{k,j}, p_k) - F_k = A_{k,j}$$

Let b_j denote the number of bad points in A_j with respect to (T, p_k), then $|G'_{k,j}| = |A_j| - b_j$. By the first condition in (3.13), we have the following upper bound:

$$\sum_{j\in[m]} b_j \le \frac{\epsilon}{\gamma}\cdot|T| \le \frac{\epsilon}{\gamma}\cdot 2mW = \frac{mW}{16\lceil\log v\rceil} \le \frac{mW}{16\log v}$$

As a result, the number of $j \in [m]$ such that b_j satisfies

$$b_j \le \frac{W}{8\log v}$$

must be at least $m/2$. We use R to denote the set of such $j \in [m]$. For each $j \in R$, we have

$$\left|G'_{k,j}\right| \ge |A_j| - \frac{W}{8\log v} \ge |A_j|\cdot\left(1 - \frac{1}{4\log v}\right) \tag{3.15}$$

as $W \le 2|A_j|$. This implies that

$$\left|\,\textsc{cone}\,(G'_{k,j}, p_k)\,\right| = s'|G'_{k,j}| \ge s'|A_j|\cdot\left(1 - \frac{1}{4\log v}\right)$$

Next for each $j \in [m]$, we let d_j denote the following integer:

$$d_j = \left|\, F_k \cap \textsc{cone}\,(G'_{k,j}, p_k)\,\right|$$

Since $\textsc{cone}\,(G'_{k,1}, p_k), \ldots, \textsc{cone}\,(G'_{k,m}, p_k)$ are pairwise disjoint and their union $\subseteq \textsc{cone}(A, p_k)$, we have

$$\sum_{j\in R} d_j \le \left|\, F_k \cap \textsc{cone}\,(A, p_k)\,\right| \le 4\epsilon|F_k| \le 4\epsilon\cdot r s'|A| = \lambda s'\cdot|A| \le \lambda s'\cdot mW$$

by the second condition of (3.13). Thus, the number of $j \in R$ such that $d_j > 4\lambda s'\cdot W$ is at most $m/4$. So at least $m/4$ many $j \in R$ satisfy both (3.15) and $d_j \le 4\lambda s' W$. For each of these j's, we first have

$$|A_{k,j}| \ge \left|\,\textsc{cone}\,(G'_{k,j}, p_k)\,\right| - d_j \ge s'|A_j|\cdot\left(1 - \frac{1}{4\log v} - 8\lambda\right) \ge s'|A_j|\cdot\left(1 - \frac{1}{2\log v}\right)$$

Moreover, it is clear that $\textsc{cone}\,(G'_{k,j}-B_j, p_k)$ and $B_{k,j}$ are disjoint. From $|G'_{k,j}\cap B_j| \le \alpha|A_j|$, we have

$$\textsc{cone}\,(G'_{k,j} - B_j, p_k) \ge s' \cdot \big(\,|G'_{k,j}| - \alpha|A_j|\,\big) \ge s'|A_j| \cdot \left(1 - \frac{1}{4\log v} - \alpha\right)$$

and thus, $|A_{k,j} - B_{k,j}|$ is at least

$$\big|\,\textsc{cone}\,(G'_{k,j} - B_j, p_k) - F_k\,\big| \ge s'|A_j| \cdot \left(1 - \frac{1}{4\log v} - \alpha\right) - d_j \ge s'|A_j| \cdot \left(1 - \alpha - \frac{1}{2\log v}\right)$$

As a result, we have

$$\begin{aligned}\frac{|A_{k,j} \cap B_{k,j}|}{|A_{k,j}|} =& \frac{|A_{k,j}| - |A_{k,j} - B_{k,j}|}{|A_{k,j}|}\\ =& 1 - \frac{|A_{k,j} - B_{k,j}|}{|A_{k,j}|}\\ \le& 1 - \left(1 - \alpha - \frac{1}{2\log v}\right)\\ =& \alpha + \frac{1}{2\log v}\end{aligned}$$

Finally we bound the number of (k, j), $k \in [r]$ and $j \in [m]$, such that $A_{k,j}$ and $B_{k,j}$ satisfy (3.14). Due to the analysis above, we know that for each $k \in [r]$, p_k is good with probability at least $1/2$; and when p_k is good, there are at least $m/4$ many $j \in [m]$ such that $A_{k,j}$ and $B_{k,j}$ satisfy (3.14). By Chernoff bound

$$\Pr\Big[\,\text{the number of good } p_k,\, k \in [r], \text{ is at least } r/4\,\Big] \ge 1 - \exp\left(-r/16\right) \ge 1 - 1/v^{64}$$

As a result, we have

$$\Pr\Big[\,\text{the number of } (k,j),\, k \in [r] \text{ and } j \in [m], \text{ that satisfy (3.14) is at least } mr/16\,\Big] \ge 1 - 1/v^{64}$$

Assuming the event above happens, we construct $\mathcal{E}$ as follows. Since

$$\left[\,\frac{s'|A_j|}{2},\, s'|A_j|\,\right] \subseteq \left[\,\frac{s'W}{4},\, s'W\,\right]$$

we can find an appropriate positive integer W' such that the number of (k, j) that satisfy both (3.14) and $W'/2 \le |A_{k,j}| \le W'$ is at least $mr/32 = m \cdot 32\lceil\log v\rceil$. For each such pair (k, j), we set $\mathcal{E}(k, j) = |A_{k,j}|$; and set $\mathcal{E}(k, j) = \text{nil}$ otherwise. It is easy to check that $\mathcal{E}$ is an (m, m', r, W')-expansion map and satisfies either condition ***a)*** or condition ***b)*** of Lemma 3.4.8, based on whether $\mathcal{E}$ matches $(\mathcal{B}, \mathbf{p})$. □

3.5.3 Interaction: Proof of Lemma 3.4.9

We start with a useful lemma for analyzing Interact.

Lemma 3.5.8. *Let $D \subset \mathcal{P}$ be a nonempty set of points with $\beta = (s/v) \cdot |D| < 1$. Then the total number of lines $L \in \mathcal{L}$ such that $|L \cap D| = 1$ is a least $(1-\beta)h|D|$.*

Proof. For each $k \in [0:s]$, let N_k denote the number of $L \in \mathcal{L}$ with $|L \cap D| = k$. By Proposition 4.12,

$$\textstyle\sum_k N_k \cdot k = h|D| \quad \text{and} \quad \sum_k N_k \cdot k(k-1) = |D| \cdot \big(|D| - 1\big) < |D|^2$$

Since $k \le k(k-1)$ for every $k \ge 2$, we have

$$\sum_{k \ge 2} N_k \cdot k \le \sum_{k \ge 2} N_k \cdot k(k-1) = \sum_k N_k \cdot k(k-1) < |D|^2$$

As a result, we have the following lower bound for N_1:

$$N_1 = h|D| - \sum_{k \ge 2} N_k \cdot k > h|D| - |D|^2 \tag{3.16}$$

On the other hand, by our assumption $\beta = (s/v) \cdot |D| < 1$ and $h \ge (v/s)$, see (3.1), we have $|D| = \beta \cdot (v/s) \le \beta h$. The lemma then follows by plugging $|D| \le \beta h$ into (3.16). □

We now prove Lemma 3.4.9.

Proof of Lemma 3.4.9. By assumption we have $(s/v) \cdot m \cdot W \ge \epsilon$, where ϵ denotes the following parameter:

$$\epsilon = \frac{1}{2^{18} \cdot \lceil \log v \rceil^2}$$

For each $j \in [m]$, let $D_j = A_j \Delta B_j$, the symmetric difference of A_j and B_j, and let $D = \cup_{j \in [m]} D_j$. Then

$$\frac{W}{4} \le (1-\alpha) \cdot |A_j| \le |D_j| \le 2W \quad \text{and} \quad |D| \ge \sum_j |A_j - B_j| \ge \sum_j \frac{|A_j|}{2} \ge \frac{m \cdot W}{4} \ge \frac{\epsilon}{4} \cdot \frac{v}{s}$$

Next we let j^* denote the smallest $j \in [m]$ such that

$$\big| \cup_{j \le j^*} D_j \big| \ge \frac{\epsilon}{4} \cdot \frac{v}{s}$$

Let $D^* = \cup_{j \le j^*} D_j$, then the way we picked j^* implies that

$$\frac{\epsilon}{4} \cdot \frac{v}{s} \le |D^*| \le \frac{\epsilon}{4} \cdot \frac{v}{s} + 2W < \frac{v}{2s}$$

Now we focus on the probability of the following event, which clearly implies (4.6).

$$\text{Event } E\colon \ \Big[\exists i \in [r] \text{ such that } |L_i \cap D^*| = 1 \Big]$$

To this end, note that $\beta = s|D^*|/v < 1/2$. Thus, by Lemma 3.5.8 we have for each $i \in [r]$:

$$\Pr\Big[|L_i \cap D^*| = 1 \Big] \ge \frac{(1-\beta)h|D^*|}{n} \ge \frac{(1-\beta)\epsilon h v}{4ns} = \frac{(1-\beta)\epsilon}{4} > \frac{\epsilon}{8}$$

By plugging in ϵ and r, we have

$$\Pr\big[E\big] \ge 1 - \Big(1 - \frac{\epsilon}{8}\Big)^r \ge 1 - \exp(\epsilon r/8) > 1 - 1/v^4$$

The lemma then follows. □

3.5.3.1 Proof of Theorem 3.4.11

In this section, we prove our main technical theorem as stated in Theorem 3.4.11.

Proof of Theorem 3.4.11. First we note that, because the total number of for-loops in **Cert** $(\mathcal{S}, f, g, p, q)$ is no more than $\lfloor \log v/2 \rfloor$, line 6, 13 or 20 can never evaluate to true when C_3 is large enough. The number of for-loops executed could be smaller than $\lfloor \log v/2 \rfloor$ because the *Expansion* and *Contraction* operations may exit the for-loop by returning nil or a shorter certificate that distinguishes p and q. By induction we can show that for each k, $(\mathcal{A}_k, \mathcal{B}_k)$ is an (m_k, W_k, α_k)-pair with

$$\alpha_k \le k \big/ \lceil \log v \rceil \le 1/2$$

So every time we apply one of the three operations: *Expansion* (Lemma 3.4.8), *Contraction* (Lemma 3.4.3) or *Interaction* (Lemma 3.4.9) in **Cert**, parameters $(m_{k-1}, W_{k-1}, \alpha_{k-1})$ of the current pair $(\mathcal{A}_{k-1}, \mathcal{B}_{k-1})$ of PDSTs must satisfy the needed assumptions, respectively.

We now prove that **Cert** can never reach line 26. To this end, we first show that when the constant C_2 is large enough, *Contraction* (line 12 in Figure 3.2) can never be executed

in two consecutive for-loops. We prove this statement by contradiction. Let us assume that contraction happens in both the kth and $(k+1)$th for-loops. Then we must have:

$$m_k = \left\lceil m_{k-1}/\lceil \log v \rceil \right\rceil, \quad W_k = O\left(\frac{s}{v} \cdot W_{k-1} \cdot \log^2 v\right) \quad \text{and} \quad \frac{s}{v} \cdot m_k \cdot W_k = \Omega\left(\frac{1}{\log^2 v}\right)$$

In addition, since $\mathcal{A}_{k-1}$ has m_{k-1} disjoint subsets of $\mathcal{P}$ and each has at least $W_{k-1}/2$ points, we have

$$\frac{m_{k-1} W_{k-1}}{2} \leq v.$$

Combining these inequalities, we get

$$s = \Omega\left(\frac{v}{m_k W_k \cdot \log^2 v}\right) = \Omega\left(\frac{v^2}{m_{k-1} \cdot s W_{k-1} \cdot \log^3 v}\right) = \Omega\left(\frac{v}{s \cdot \log^3 v}\right) = \Omega\left(\frac{\sqrt{n}}{\log^3 v}\right)$$

Combining it with $h = \Theta(\sqrt{n})$, we immediately have that

$$\frac{n}{sh} = O(\log^3 v)$$

which cannot happen when C_2 is set to be sufficiently large, and hence we reach a contradiction.

Since **Cert** never performs two consecutive *Contraction* operations in the for-loop, we have

$$\frac{m_{k+1}}{m_{k-1}} \geq \frac{32\lceil \log v \rceil}{\lceil \log v \rceil} = 32$$

Let $K = \lfloor \log v/2 \rfloor$, the total number of for-loops. If line 26 is reached, then by the end m_K is at least

$$32^{K/2} \text{ if } K \text{ is even; and } \frac{32^{(K-1)/2}}{\lceil \log v \rceil} \text{ if } K \text{ is odd}$$

which is larger than v in both cases. But this cannot happen because the union of all sets in an (m, W)-PDST has at least m points, and hence $m_K \leq v$. Now we have shown that line 26 is never reached.

Therefore, for any two points $p, q \in \mathcal{P}$, $\mathbf{Cert}(\mathcal{S}, f, g, p, q)$ returns nil only if line $9, 16$ or 23 evaluates to true. By Lemma 3.4.3, Lemma 3.4.8 and Lemma 3.4.9, if f and g are sampled uniformly at random then the probability that line $9, 16$ or 23 never evaluates to true in any of the $\lfloor \log v/2 \rfloor$ for-loops is at least

$$\left(1 - \frac{1}{v^3}\right)^{\lfloor \log v/2 \rfloor} = 1 - O\left(\frac{\log v}{v^3}\right)$$

Since there are $O(v^2)$ many pairs of $p, q \in \mathcal{P}$, the lemma follows using the union bound. □

Chapter 4

Isomorphism of strongly regular graphs

In this chapter we analyze the structure and isomorphism of strongly regular graphs.

Definition 4.0.1. *A* strongly regular(SR for short) graph *with parameters* (n, k, λ, μ) *is a* k*-regular graph on* n *vertices, in which every two adjacent vertices have* λ *common neighbors and every two non-adjacent vertices have* μ *common neighbors.*

We present an $\exp(\widetilde{O}(n^{1/5}))$ time algorithm for isomorphism testing of strongly regular graphs. We also show that every non trivial strongly regular graph has at most $\exp(\widetilde{O}(n^{9/37}))$ automorphisms.

4.1 Strongly regular graphs

In this section, we present some basic facts about strongly regular graph.

All the disconnected SR graphs are disjoint unions of cliques of the same size. We refer to these graphs and their complements as *trivial* SR graphs. The following facts about non-trivial SR graphs can be easily seen from the definition.

Proposition 4.1.1. *Let* G *be a non-trivial SR graph with parameters* (n, k, λ, μ)*. Then,*

1. G *has diameter* 2*.*

2. $k \geq \sqrt{n-1}$.

3. $\mu(n-k-1) = k(k-\lambda-1)$.

4. *The adjacency matrix of* G *has exact three distinct eigenvalues,* $k > r \geq 0 \geq s$.

Without loss of generality, we also assume throughout the chapter that k satisfies $k \leq (n-1)/2$ since the complement of a SR graph is SR.

Given a graph $G = (V, E)$, the line-graph $L(G)$ has the vertex set E, and two vertices in $L(G)$ are adjacent if the corresponding edges in G share a vertex. It is a folklore that for $L(G)$ to be a non-trivial SR graph, G must be either a complete graph, a complete bipartite graph or the 5-cycle graph. These graphs and their complements are referred to as *graphic* SR graphs. The following result was central to Spielman's work and remains central to ours.

Theorem 4.1.2 (Neumaier)**.** *Let* G *be a* non-trivial *and* non-graphic *SR graph with parameters* (n, k, λ, μ) *and eigenvalues* $k > r > s$*. Then, at least one of the following conditions must hold:*

(S) $\mu = s^2$ *and* G *is a* Steiner graph *derived from a* Steiner 2-design*;*

(L) $\mu = s(s+1)$ *and* G *is a* Latin square graph *derived from an* s-net*;*

(F) G *is a conference graph* (*and thus,* $k = (n-1)/2$*,* $\mu = (n-1)/4$*, and* $\lambda = \mu - 1$)*;*

(C) G *satisfies* Neumaier's claw bound*:*

$$r \leq \max\left\{ 2(-s-1)(\mu+1+s) + s,\ \frac{s(s+1)(\mu+1)}{2} - 1 \right\}. \tag{4.1}$$

Steiner and Latin square graphs are both defined by a *finite geometry* that consists of a set of points and a set of "lines" each of which is itself a subset of points. As we will not examine these *geometric* SR graphs, we refer interested readers to [Neumaier, 1979; Spielman, 1996; Miller, 1978; Babai and Wilmes, 2013; Chen *et al.*, 2013] for their definitions.

Lemma 4.1.3. *Let G be an SR graph with parameters (n, k, λ, μ) and eigenvalues $k > r > 0 > s$. If G satisfies Neumaier's claw bound and $k = o(n)$, then we have*

$$\mu = \frac{k^2}{n} \cdot \big(1 \pm o(1)\big) \quad \textit{and} \quad \lambda = O\left(\frac{k^{4/3}}{n^{1/3}}\right).$$

Proof. Following the proof of Corollary 9 of [Spielman, 1996] we have $\mu = o(k)$ and $r = O(k^{2/3}\mu^{1/3})$. By Part (b) of Proposition 2 of [Spielman, 1996], we have the following bound for λ:

$$\lambda = \mu + r + s \leq \mu + r \leq o\big(k^{2/3}\mu^{1/3}\big) + O(k^{2/3}\mu^{1/3}) = O(k^{2/3}\mu^{1/3}).$$

This implies that $\lambda = o(k)$. On the other hand, from $\mu(n-k-1) = k(k-\lambda-1)$ we have

$$\mu = \frac{k(k-\lambda-1)}{n-k-1} = \frac{k^2(1-o(1))}{n(1-o(1))} = \frac{k^2}{n} \cdot \big(1 \pm o(1)\big)$$

Plugging this into $\lambda = O(k^{2/3}\mu^{1/3})$, the lemma follows directly. □

We also have

Theorem 4.1.4 ([Babai and Wilmes, 2015])**.** *Let G be a non-trivial SR graph. Then*

$$\lambda = O(k^{3/2}n^{-1/2} + n^{1/2}). \tag{4.2}$$

Consequently, for $k = \Omega(n^{2/3})$ we have $\lambda = O(\sqrt{k\mu})$.

4.2 Isomorphism of SR graphs

The class of strongly regular graphs, while not believed to be GI-complete, has long been identified as a hard case for GI (cf. [Read and Corneil, 1977]).

In [Babai, 1981b], Babai proved

Theorem 4.2.1. *Every SR graph G with n vertices and $k \leq (n-1)/2$ has a set of $\widetilde{O}(n/k)$ vertices whose individualization completely splits the graph under naive vertex refinement.*

This gives an isomorphism testing algorithm for strongly regular graphs with running time $\exp(\tilde{O}(\sqrt{n}))$. Spielman [Spielman, 1996] then showed

Theorem 4.2.2. *Every SR graph with $k = o(n^{2/3})$ and second eigenvalue $r = o(k)$ has a set of $\tilde{O}(\sqrt{n/k})$ vertices whose individualization completely splits the graph under naive vertex refinement.*

This implies an isomorphism testing algorithm for strongly regular graphs with running time $\exp(\widetilde{O}(n^{1/3}))$. In this section, we further improve this bound to $\exp(\widetilde{O}(n^{1/5}))$.

Theorem 4.2.3. *Let G be a SR graph with n vertices. Then a canonical form for G can be computed in time* $\exp(\widetilde{O}(n^{1/5}))$.

Theorem 4.2.3 is obtained by applying the following theorem.

Theorem 4.2.4. *Let G be a SR graph with n vertices and degree $k \leq (n-1)/2$. Then a canonical form for G can be computed in time*

(a) $\exp(\widetilde{O}(1 + k^2/n))$;

(b) $\exp(\widetilde{O}(\sqrt{n/k}))$;

(c) $\exp(\widetilde{O}(n/k))$.

Theorem 4.2.3 is obtained by applying (a) for $k \leq n^{3/5}$, part (b) for $n^{3/5} \leq k \leq n^{3/4}$ and part (c) for $n^{3/4} \leq k \leq (n-1)/2$.

In this section, we present a proof of Theorem 4.2.4 (b). For the proof of (a), (c) and an alternative proof of Theorem 4.2.4, see [Babai *et al.*, 2013] and [Wilmes, 2016]. More precisely, we prove the following result.

Theorem 4.2.5. *Fix any constant $\varepsilon > 0$. Every non-trivial vertex-colored SR graph satisfying $k = \Omega(n^{2/3})$ and $k = O(n^{1-\varepsilon})$ is completely split by $\widetilde{O}((n/k)^{1/2})$ vertices under classical Weisfeiler-Leman refinement.*

We prove Theorem 4.2.5 by presenting a canonical pairwise distinguisher for vertices of a strongly regular graph.

We will construct two technical algorithms **Cert** and **Test** for SR graphs. The first algorithm **Cert** (G, f, u, v) iteratively applies the operation Partition to build two multi-stage branching structures, one for u and one for v, and uses another operation Interact

to certify the pairwise distinctness at the final stage. When it succeeds, **Cert** returns a certificate M that is essentially a sketch of the structure built for u. The second algorithm **Test** (G, f, M, u) then applies the same operations Partition and Interact, trying to build a multi-stage structure for u that matches the description given in M. It outputs 1 if it succeeds; and 0 if it fails.

The **Cert** algorithm aims at distinguishing two vertices $u \neq v$ in G. We simultaneously grow two sequences of such bipartite structures, one from u and one from v, with the assistance of a small number of individualized vertices (referred to as *seeds*) sampled independently and uniformly at random. These bipartite structures are grown in a canonical fashion (i.e., if ϕ is an isomorphism from G to G', then bipartite structures grown from u and $\phi(u)$ will be the same under the isomorphism ϕ).At each step, the pair of bipartite structures grown so far have the following property. Either their "interactions" with a small number of random seeds can introduce the desired asymmetry between u and v with high probability, or their "interactions" with a small number of random seeds likely produce another pair of bipartite structures with measurable progress towards the former case. Below we explain in more details these "interactions".

For a vertex u, we focus on the induced bipartite subgraph between $N(u)$ and $V \setminus N^+(u)$. Specifically, we focus on a family of *bipartite systems*, each consisting of a sequence of induced bipartite subgraphs $((A_1, B_1), \ldots, (A_\gamma, B_\gamma))$, where the A_i are disjoint subsets of $N(u)$ and $B_i \subseteq V \setminus N^+(u)$. For our construction and analysis, in addition to demanding that every bipartite graph induced by A_i, B_i is dense enough, we further require that

- the sizes of all the A_i be within a factor of 2 of each other,
- all degrees involved by vertices in $\bigcup_i B_i$ be within a factor of 2 of each other,
- the numbers of B_i to which each vertex in $\bigcup_i B_i$ belongs be within a constant factor of each other.

(See Definition 4.2.7 for more details.)

With these strong "regularity" conditions, we have the following property: Suppose $((A_i, B_i))$ and $((A'_i, B'_i))$ are a pair of bipartite systems built from u, v, respectively. If $|A_i \cap A'_i| = o(|A_i|)$ and $|A_i| = |A'_i|$ is smaller than $k/\max(\lambda, \mu)$, then we have $|B_i \cap B'_i| =$

$o(|B_i|)$ (and otherwise we are already done in distinguishing u and v). Thus, if $|\bigcup_i B_i|$ is very close to n, then the interaction of a small number of random seeds w with B_i and B'_i is likely to produce the asymmetry that we aim for: for some i, $w \in B_i$ but $w \notin B'_i$ (since these bipartite structures are built in a canonical fashion).

The two initial bipartite systems for u and v are simply $((N(u), V \setminus N^+(u))$ and $((N(v), V \setminus N^+(v))$ which do not meet the size condition above as $|N(u)| = |N(v)| = k$ is too large. To make progress, we draw a small number of fresh random seeds and use the following process to *partition* a bipartite system $((A_i, B_i))$ to $((A^*_j, B^*_j))$. For a seed z, if $z \in B_i$ for some i, then we extract a new induced bipartite graph (A^*, B^*), where A^* contains all the vertices of A_i that are also neighbors of z, and B^* contains all neighbors of A^* in B_i. We collect all such new induced bipartite graphs and then "*clean them up*" to make sure the new bipartite system satisfy the desired regularity conditions (i–iii) again. Our two goals are to ensure that (1) the union of the B-part of the new bipartite system still contains almost all vertices in $V \setminus N^+(u)$ and (2) the A-part of the new system is smaller than the old one by a factor of $O(n^{-\Omega(1)})$. Given these two properties, a constant number of steps is sufficient to obtain the desired bipartite system for u.

While the partition operation is intuitively simple, the greatest challenge for us is to make sure that the new bipartite system remains highly regular (i.e., still satisfies all conditions (i–iii); also see Definition 4.2.7). For this purpose, we need to first extract from the old bipartite system $((A_i, B_i))$ layers of structures that satisfy much stronger regularity conditions than (i–iii), as a preparation for the partition operation. This involves careful definitions of special vertices and pairs (A_i, B_i) satisfying those stronger regularity conditions, as well as lemmas that show abundance of such objects. After using random seeds to further partition $((A_i, B_i))$, as described above, we apply a carefully designed cleaning up procedure which only makes relatively minor changes to $((A^*_j, B^*_j))$ but produces at the end a bipartite system that satisfies again all regularity conditions (i–iii).

Given the partition operation we show that our construction of bipartite systems is canonical, and that $\widetilde{O}((n/k)^{1/2})$ random seeds are sufficient to distinguish all pairs of vertices with high probability. From these results, we derive a canonical pairwise distinguisher for the graph, and show that a seeding set of size $\widetilde{O}((n/k)^{1/2})$ is enough to distinguish every pair

of vertices. For technical reasons, our analysis works only for SR graphs with $k = O(n^{1-\varepsilon})$, for any arbitrary constant $\varepsilon > 0$.

4.2.1 Bipartite systems

Fix any constant $\varepsilon > 0$. Let $G = (V, E)$ be a SR graph with parameters (n, k, λ, μ). We will only require the assumptions $k = \Omega(n^{2/3})$ and $k = O(n^{1-\varepsilon})$ in Section 4.2.4. We note that it then follows that $\lambda = O(\sqrt{k\mu})$ by Theorem 4.1.4. In the remainder of Section 4.2.1, we require only the weaker assumptions that $k = o(n)$ (and hence $\mu = o(k)$ by Lemma 4.1.3), and $\lambda = o(k)$.

We need some notation. Given $p \notin A \subseteq V$, we let $E(p, A)$ denote the set of edges $(p, q) \in E$ with $q \in A$. Given two disjoint sets $A, B \subseteq V$, let $E(A, B)$ denote the set of edges between A and B. Sometimes we use $E(A, B)$ to denote the bipartite graph induced by A and B when it is clear from the context. Given $u \in V$, we let H_u denote the bipartite subgraph of G induced by $N(u)$ and $V \setminus N^+(u)$. Let $\alpha = \lfloor \log n \rfloor$.

In Section 4.2.1, u is a fixed vertex so we suppress the subscript and denote H_u by H. Our algorithm is based on the following bipartite structures.

Definition 4.2.6 (Bipartite Systems)**.** We call $\mathcal{S} = ((A_i, B_i) : i \in [\gamma])$ a *bipartite system* in H of size $\gamma \geq 1$ if $\mathcal{S}$ satisfies the following two conditions:

1. $A_i \subseteq N(u)$ and $B_i \subseteq V \setminus N^+(u)$ are nonempty for all $i \in [\gamma]$, and
2. $A_i \cap A_j = \emptyset$ for all $i \neq j \in [\gamma]$. (The B_i are not necessarily pairwise disjoint.)

Given $\mathcal{S} = ((A_i, B_i) : i \in [\gamma])$, we let $A = \bigcup_i A_i$ and $B = \bigcup_i B_i$. Let $M(p, \mathcal{S}) = \{i : p \in B_i\}$ denote the number of times p appears in B_i, for some $p \in B$.

We will use the following "highly regular" bipartite systems.

Definition 4.2.7. Let γ, m, h, t be positive integers, and $\rho \in (0, 1]$. We call $\mathcal{S} = ((A_i, B_i) : i \in [\gamma])$ a (γ, m, h, t, ρ)-*bipartite system* in H if it is a bipartite system of size γ and satisfies the following conditions:

1. For all $i \in [\gamma]$, we have $m/2 \leq |A_i| \leq m$.
2. For all $i \in [\gamma]$ and $p \in B_i$, we have $h/2 \leq |E(p, A_i)| = |N(p) \cap A_i| \leq h$.

3. For all $i \in [\gamma]$, $|E(A_i, B_i)| \geq \rho mk$ (also bounded from above trivially by mk).

4. For all $p \in B$, we have $t/8 \leq |M(p, \mathcal{S})| \leq t$.

By definition, $((N(u), V \setminus N^+(u)))$ is a trivial $(1, k, \mu, 1, 1 - o(1))$-bipartite system in H initially (since $\lambda = o(k)$).

The following lemma follows directly from definitions above.

Lemma 4.2.8. *If $\mathcal{S} = ((A_i, B_i) : i \in [\gamma])$ is a (γ, m, h, t, ρ)-bipartite system, then*

$$\frac{\rho mk}{h} \leq |B_i| \leq \frac{2mk}{h} \quad \text{and} \quad \frac{\rho mk\gamma}{ht} \leq |B| \leq \frac{16mk\gamma}{ht}.$$

The next lemma gives us a very useful upper bound on h.

Lemma 4.2.9. *Either $h = O(1)$ or $h = O\left(m \cdot \max(\lambda, \mu)/(\rho k)\right)$.*

Proof. Suppose $h \geq 4$. Fix an $i \in [\gamma]$, and we count the number of triples (p, a, b) such that $p \in B_i$, $a \neq b \in A_i$, and $(p, a), (p, b) \in E$. Because $|E(p, A_i)| \geq h/2 \geq 2$, by picking p first, this number is at least $|B_i| \cdot \Theta(h^2) \geq (\rho mk/h) \cdot \Theta(h^2) = \Theta(\rho mkh)$. On the other hand, by picking a and b first, this number is at most $m^2 \cdot \max(\lambda, \mu)$. This finishes the proof and the lemma follows. □

4.2.2 Partition

Let $\mathcal{S} = ((A_i, B_i) : i \in [\gamma])$ be a (γ, m, h, t, ρ)-bipartite system in H. For each $i \in [\gamma]$ and $p \in B_i$, we say p is *good* in B_i if the number of edges between B_i and neighbors of p in A_i is large:

$$\left|E(A_i \cap N(p), B_i)\right| \geq \rho^2 hk/4.$$

The following lemma shows that many vertices in each B_i are good (in B_i).

Lemma 4.2.10. *Let $\mathcal{S} = ((A_i, B_i) : i \in [\gamma])$ be a (γ, m, h, t, ρ)-bipartite system in H. Then for each $i \in [\gamma]$, the number of good vertices in B_i is at least $\rho^2 mk/(2h)$.*

Proof. Fix an $i \in [\gamma]$. We now count the number L of triples (p, a, b) such that $p, b \in B_i$ (though they are not necessarily distinct) and $a \in A_i$, with $(p, a), (a, b) \in E$. If we pick $a \in A_i$ first, then both p and b have $|E(a, B_i)|$ choices. By Cauchy–Schwarz,

$$L = \sum_{a \in A_i} |E(a, B_i)|^2 \geq \frac{\left(\sum_{a \in A_i} |E(a, B_i)|\right)^2}{|A_i|} = \frac{|E(A_i, B_i)|^2}{|A_i|}.$$

Using $|E(A_i, B_i)| \geq \rho mk$ and $|A_i| \leq m$, we have $L \geq \rho^2 mk^2$.

On the other hand, if we pick $p \in B_i$ first, then we have

$$\sum_{p \in B_i} \left|E(A_i \cap N(p), B_i)\right| = L \geq \rho^2 mk^2.$$

Let T denote the number of good vertices in B_i. As $|E(A_i \cap N(p), B_i)| \leq hk$, we have

$$\frac{2mk}{h} \cdot \frac{\rho^2 hk}{4} + T \cdot hk \geq \rho^2 mk^2,$$

which implies that $T \geq \rho^2 mk/(2h)$, and the lemma is proven. □

Next, for each $i \in [\gamma]$, we introduce the following function $F_i(p, q)$ over $p, q \in B_i$ (note that p, q here are not necessarily distinct): $F_i(p, q) = |A_i \cap N(p) \cap N(q)|$, i.e., $F_i(p, q)$ is the degree of q in $E(A_i \cap N(p), B_i)$, the bipartite graph induced by $A_i \cap N(p)$ and B_i. For a good vertex p in B_i, by definition we have

$$\sum_{q \in B_i} F_i(p, q) = \left|E(A_i \cap N(p), B_i)\right| \geq \frac{\rho^2 hk}{4}.$$

We define the *type* of a good vertex p in B_i, denoted $\textsc{Type}_i(p)$, as a positive power d of 2 that maximizes:

$$\sum_{\substack{q \in B_i \\ d \leq F_i(p,q) < 2d}} F_i(p, q),$$

with tie-breaking done by picking the smallest such d. (This will be the tie-breaking rule used in this section by default.) Equivalently, we put vertices of B_i in "buckets" of exponentially increasing sizes with respect to their degrees in $E(A_i \cap N(p), B_i)$; the type of p is then the "bucket" with the largest total degree. Note that type of p may vary in the different sets B_i (as suggested in the subscript i in $\textsc{Type}_i$).

By an averaging argument, if p is a type-d good vertex in B_i, then we have

$$\sum_{\substack{q \in B_i \\ d \leq F_i(p,q) < 2d}} F_i(p, q) \geq \frac{\rho^2 hk}{4\alpha},$$

since $F_i(p, q) \leq \mu = o(n)$ so the number of buckets is $o(\log n) = o(\alpha)$.

We say the *type of a bipartite system* is d, a positive power of 2, if d maximizes the number of pairs (i, p) such that p is good in B_i and $\textsc{Type}_i(p) = d$. So the type of a

bipartite system $\mathcal{S}$ is the most popular type among all such pairs. By Lemma 4.2.10 and an averaging argument, the number of (i, p) such that $\textsc{Type}_i(p) = d$ is at least

$$\gamma \cdot \frac{\rho^2 mk}{2h} \cdot \frac{1}{\alpha} = \frac{\rho^2 mk\gamma}{2\alpha h}.$$

Given d, the type of $\mathcal{S}$, we focus on vertices in B that appear in many pairs (i, p) with $\textsc{Type}_i(p) = d$. We say $p \in B$ is a *good vertex in the system* $\mathcal{S}$ if it appears in at least $\rho^2 t/(64\alpha)$ many such pairs, i.e., p is a good type-d vertex in at least $\rho^2 t/(64\alpha)$ of the B_i. The next lemma shows that there are many good vertices in $\mathcal{S}$.

Lemma 4.2.11. *The number of good vertices in $\mathcal{S}$ is at least $\rho^2 mk\gamma/(4\alpha ht)$.*

Proof. Let T denote the number of good vertices in $\mathcal{S}$. Then we have

$$|B| \cdot \frac{\rho^2 t}{64\alpha} + T \cdot t \geq \frac{\rho^2 mk\gamma}{2\alpha h}.$$

The lemma follows by the upper bound of $|B|$ in Lemma 4.2.8 and solving for T. □

From now on, we use d to denote the type of $\mathcal{S}$. We focus on a good vertex p in $\mathcal{S}$ and classify it further according to how it is connected with vertices in B. Given a vertex $q \in B$ (p, q here are not necessarily distinct), we say p and q have *a connection of strength* $s \geq 0$, denoted by $\textsc{Str}(p, q)$, if there are s indices $i \in [\gamma]$ such that $F_i(p, q)$ satisfies $d \leq F_i(p, q) \leq 2d - 1$. Since p is a good vertex in $\mathcal{S}$, by definition we have

$$\sum_{q \in B} \textsc{Str}(p, q) \geq \frac{\rho^2 t}{64\alpha} \cdot \frac{\rho^2 hk}{4\alpha} \cdot \frac{1}{2d} = \frac{\rho^4 hkt}{2^9 \alpha^2 d}.$$

We then put q in "buckets" of exponentially increasing sizes according to the strength $\textsc{Str}(p, q)$ between p and q and refer to the "bucket" with the largest total strength as the *strength* of p. More formally, we say the strength of a good vertex p in a bipartite system $\mathcal{S}$ is a positive power s of 2 that maximizes:

$$\sum_{\substack{q \in B \\ s \leq \textsc{Str}(p,q) < 2s}} \textsc{Str}(p, q).$$

By an averaging argument the sum above is at least $\rho^4 hkt/(2^9 \alpha^3 d)$.

Finally, the *strength of a bipartite system* $\mathcal{S}$ is a positive power s of 2 that maximizes the total number of good vertices of strength s in the bipartite system (i.e., the most popular

strength s among all good vertices). For the rest of the section, we let d denote the type and s denote the strength of the system $\mathcal{S}$ being considered, both of which are positive powers of 2 (including 1).

The next definition will play a crucial role in further refining a bipartite system using a small set of random seeds as discussed in the next subsection.

Definition 4.2.12. Given a (γ, m, h, t, ρ)-bipartite system $\mathcal{S}$ of type d and strength s, a *buzzer* is a good vertex $p \in B$ (i.e., p is a good, type-d vertex in at least $\rho^2 t/(64\alpha)$ of the B_i) of strength s. We call $q \in B$ a *receiver* for a buzzer p if $s \leq \textsc{Str}(p, q) < 2s$. We call $i \in [\gamma]$ a *dispatcher* for a buzzer $p \in B$ if the number of receivers $q \in B$ for p such that $d \leq F_i(p, q) < 2d$ is at least $\rho^4 hk/(2^{10}\alpha^3 d)$.

In the last two definitions, p and q are not necessarily distinct. Note that $q \in B_i$ being a receiver for p does not necessarily imply the $d \leq F_i(p, q) < 2d$: it implies that $d \leq F_j(p, q) < 2d$ for $[s : 2s - 1]$ many j but not necessarily every i with $q \in B_i$.

Our plan is to use a small set of random seeds (vertices) $y_1, \ldots, y_\theta$ to partition $\mathcal{S}$, and define a new bipartite system $\mathcal{S}' = \{(C_k, D_k)\}$ with *smaller* sets C_k: Roughly speaking, each C_k is the intersection of A_i and $N(y_j)$, for some i, j, and D_k is a subset of $N(C_k) \cap B_i$. The challenge, however, is to make sure that $\mathcal{S}'$ is again a highly regular bipartite system (in the sense of Definition 4.2.8) but with measurable progress on its parameter m. For these purposes, buzzers will serve as candidates for y_j; given a buzzer y_j, we will add one pair (C_k, D_k) to $\mathcal{S}'$ by setting $C_k = A_i \cap N(y_j)$ for each dispatcher i of y_j, and setting D_k to be the set of receivers of y_j in B_i. While $\mathcal{S}'$ violates many conditions of Definition 4.2.8, properties of these objects (either from their definitions or lemmas below) allow us to clean up and regularize $\mathcal{S}'$ to obtain a bipartite system that fits Definition 4.2.8 and has a smaller parameter m as desired.

We have the following corollary from Lemma 4.2.11 and the definition of buzzers.

Corollary 4.2.13. *The number of buzzers is* $\Omega(\rho^2 mk\gamma/(\alpha^2 ht))$.

We also bound the number of dispatchers for each buzzer as follows.

Lemma 4.2.14. *Each buzzer* $p \in B$ *has at least* $\Omega(\rho^4 t/\alpha^3)$ *dispatchers* $i \in [\gamma]$.

Proof. Since p is a buzzer in $\mathcal{S}$ (good vertex in $\mathcal{S}$ of strength s), we have

$$\sum_{\substack{q\in B\\ s\le \textsc{Str}(p,q)<2s}} \textsc{Str}(p,q) \ge \frac{\rho^4 hkt}{2^9\alpha^3 d}.$$

Let T denote the number of dispatchers for p. Then we have

$$t\cdot\frac{\rho^4 hk}{2^{10}\alpha^3 d} + T\cdot\frac{hk}{d} \ge \frac{\rho^4 hkt}{2^9\alpha^3 d}.$$

The lemma follows by solving the inequality for T. □

We end this subsection with a lemma concerning the type d and strength s of a bipartite system $\mathcal{S}$.

Lemma 4.2.15. *Either $d = s = 1$, or we have $ds = O(\alpha^3 ht\cdot\max(\lambda,\mu)/(\rho^4 k))$.*

Proof. Assume $ds > 1$. Fix a buzzer p with Q being the set of its receivers. Then

$$\Theta(|Q|\cdot s) = \sum_{q\in Q}\textsc{Str}(p,q) = \Omega\left(\frac{\rho^4 hkt}{\alpha^3 d}\right) \;\Rightarrow\; |Q|\cdot ds = \Omega\left(\frac{\rho^4 hkt}{\alpha^3}\right). \tag{4.3}$$

On the other hand, we count the number of triples (a,b,q) such that $a\neq b\in A$, q is a receiver of p, and satisfies $(a,p),(a,q),(b,p),(b,q)\in E$. By picking $a,b\in A$ first, we see that the number of such triples is $\le (ht)^2\max(\lambda,\mu)$. On the other hand, using the $|Q|$ receivers, we can find at least $|Q|\cdot ds(ds-1) = |Q|\cdot\Omega((ds)^2)$ such triples (as $ds>1$). Thus, $|Q|\cdot\Omega((ds)^2)\le (ht)^2\max(\lambda,\mu)$. The lemma follows by combining this inequality and (4.3). □

Let $\mathcal{S} = ((A_i,\ B_i) : i\in[\gamma])$ be a (γ,m,h,t,ρ)-bipartite system in H of type d and strength s, and let $y_1,\ldots,y_\theta$ denote a sequence of $\theta\ge 1$ vertices sampled independently and uniformly at random. In this subsection, we use $y_1,\ldots,y_\theta$ to further partition $\mathcal{S}$ and construct a new bipartite system with a *smaller* parameter m.

Let $R = mk\theta/(hn)$ (which can be viewed as the expected number of $y_1,\ldots,y_\theta$ in a set B_i). We always assume in this section that the two parameters θ and R satisfy

$$\theta\le O\big(\rho^{10}n/(\alpha^{10}k)\big) \quad\text{and}\quad R\ge \alpha^8/\rho^6, \tag{4.4}$$

which is guaranteed whenever we apply the partition operation later in §4.2.3.

Given $y_1, \ldots, y_\theta$, we let T denote the set of all pairs (i, j), $i \in [\gamma]$ and $j \in [\theta]$, such that y_j is a buzzer and i is a dispatcher for y_j. The following lemma bounds $|T|$.

Lemma 4.2.16. *With probability at least* $1 - \exp(-\Omega(\alpha^2))$, $|T| \geq \Omega(\rho^6 \gamma R/\alpha^5)$.

Proof. By Corollary 4.2.13, each y_j is a buzzer with probability $\Omega(\rho^2 mk\gamma/(\alpha^2 nht))$. As a result the expected number of buzzers sampled is $\Omega(\rho^2\gamma R/(\alpha^2 t)) = \omega(\alpha^2)$ using (4.4). By the Chernoff bound, with probability $1 - \exp(-\Omega(\alpha^2))$, number of buzzers sampled in $y_1, \ldots, y_\theta$ is $\Omega(\rho^2\gamma R/(\alpha^2 t))$. The lemma follows from Lemma 4.2.14. □

We prove an upper bound for the number of times each $i \in [\gamma]$ appears in T.

Lemma 4.2.17. *Fix an* $i \in [\gamma]$. *With probability at least* $1 - \exp(-\Omega(\alpha^2))$, *we have* $|\{j : (i, j) \in T\}| \leq O(R)$.

Proof. Since $|B_i| = O(mk/h)$, the expected number of samples $y_1, \ldots, y_\theta$ in B_i is $O(R)$. The lemma follows directly from $R = \omega(\alpha^2)$ and the Chernoff bound. □

Next, for each pair $(i, j) \in T$, we set

$$C_{i,j} = \big(A_i \cap N(y_j)\big) \setminus \left(\bigcup_{\substack{j' < j \\ (i,j') \in T}} \big(A_i \cap N(y_{j'})\big) \right).$$

Since the A_i are pairwise disjoint, all the sets $C_{i,j}$, $(i, j) \in T$, are pairwise disjoint as well. Given a pair $(i, j) \in T$ we say that $C_{i,j}$ *overlaps* if $|C_{i,j}| < |A_i \cap N(y_j)| - \rho^4 h/\alpha^4$. Let $T^* \subseteq T$ denote the set of pairs $(i, j) \in T$ such that $C_{i,j}$ does not overlap.

We will only keep those $C_{i,j}$ that do not overlap. To obtain an upper bound for the number of those $C_{i,j}$ that overlap, for each $i \in [\gamma]$ we let J_i denote the set of the smallest $O(R)$ indices ℓ such that $(i, \ell) \in T$. (Here the constant hidden in $O(R)$ is chosen to be the same as that in Lemma 4.2.17.) We then let

$$C'_{i,j} = \big(A_i \cap N(y_j)\big) \setminus \left(\bigcup_{j > j' \in J_i} \big(A_i \cap N(y_{j'})\big) \right).$$

By definition, we have $C'_{i,j} = C_{i,j}$ when Lemma 4.2.17 holds (which happens with high probability). Similarly, we say $C'_{i,j}$ overlaps if $|C'_{i,j}| < |A_i \cap N(y_j)| - \rho^4 h/\alpha^4$.

Lemma 4.2.18. *Let $i \in [\gamma]$. With probability at least $1 - \exp(-\Omega(\alpha^2))$, the number of $j \in [\theta]$ such that $(i,j) \in T$ and $C'_{i,j}$ overlaps is at most $O(\rho^6 R/\alpha^6)$.*

Proof. We examine the samples $y_1, \ldots, y_\theta$ one by one. For each j, the number of vertices that have $\rho^4 h/\alpha^4$ edges to the union of $A_i \cap N(y_{j'})$, $j' < j$ and $j' \in J_i$, is

$$O(R) \cdot hk \cdot \alpha^4/(\rho^4 h) = O(\alpha^4 kR/\rho^4).$$

Thus, the probability that $C'_{i,j}$ overlaps is $O(\alpha^4 kR/(\rho^4 n))$, and the expected number of $j \in [\theta]$ such that $C'_{i,j}$ overlaps is (using the assumption on θ):

$$O\big(\alpha^4 kR\theta/(\rho^4 n)\big) = O\big(\rho^6 R/\alpha^6\big).$$

Using the (generalized) Chernoff bound, with probability at least $1 - \Omega(\rho^6 R/\alpha^6)$ the number of j such that $C'_{i,j}$ overlaps is $O(\rho^6 R/\alpha^6)$. The lemma then follows from the assumption on R in (4.4). □

To summarize, we get the following corollary from Lemma 4.2.16, 4.2.17 and 4.2.18:

Corollary 4.2.19. *With probability $1 - \exp(-\Omega(\alpha^2))$, $|T^*| = \Omega(\rho^6 \gamma R/\alpha^5)$.*

Below we assume that the event in Corollary 4.2.19 happens. For each $(i,j) \in T^*$, y_j is a buzzer and i is a dispatcher for y_j. By definition there are at least $\Omega(\rho^4 hk/(\alpha^3 d))$ receivers q of y_j that satisfy $q \in B_i$ and $d \le F_i(y_j, q) < 2d$, and we use $D_{i,j}$ to denote the set of all such q. Since $|C_{i,j}| \ge |A_i \cap N(y_j)| - \rho^4 h/\alpha^4$ and the $C_{i,j}$ are pairwise disjoint, we have the following lemma.

Lemma 4.2.20. *For each pair $(i,j) \in T^*$, we have $|E(C_{i,j}, D_{i,j})| = \Omega(\rho^4 hk/\alpha^3)$ and $|E(q, C_{i,j})| < 2d$ for all $q \in D_{i,j}$.*

Before moving on, we record a lemma that will be helpful later when comparing the two sequences of bipartite structures built from two vertices u and v.

Lemma 4.2.21. *Let W be a subset of A. Then with probability $1 - \exp(-\Omega(\alpha^2))$,*

$$\sum_{(i,j) \in T^*} |W \cap C_{i,j}| \le \frac{|W| k\theta}{n} + \alpha\sqrt{\theta} ht.$$

Proof. From the definition of bipartite systems, $\sum_i |W \cap C_{i,j}|$ is bounded from above by $\min(|N(y_j) \cap W|, ht)$ with probability 1. Because the latter has expectation at most $|W|k/n$, is independent of each other, and is bounded from above by ht with probability 1. The lemma then follows from Hoeffding bound:

$$\mathsf{Pr}\left[\sum X_j - \frac{|W|k\theta}{n} \geq \alpha\sqrt{\theta}ht\right] \leq \exp\left(-\frac{2(\alpha\sqrt{\theta}ht)^2}{\theta(ht)^2}\right) = \exp(-2\alpha^2).$$

□

Bounding the number of occurrences

Let D denote the union of the $D_{i,j}$ over all $(i,j) \in T^*$. We would like to prove an upper bound on the number of occurrences of a vertex p in $D_{i,j}$, $(i,j) \in T^*$. Let $L = thk\theta/(dsn)$ and $\alpha = \lfloor \log n \rfloor$.

Lemma 4.2.22. *With probability* $1 - \exp(-\Omega(\alpha^2))$, *each vertex* $p \in D$ *appears in at most* $O(s \cdot \max(L, \alpha^2))$ *many of the* $D_{i,j}$.

Proof. Let $p \in B$. Let Q denote the set of buzzers q such that p is a receiver for q. We have $|Q| \leq O(thk/(ds))$. Note that the number of occurrences of p in the $D_{i,j}$ can be easily bounded from above by s times the number of vertices of Q sampled in seeds $y_1, \ldots, y_\theta$. The latter has expectation $O(thk\theta/(dsn)) = O(L)$. By the Chernoff bound, with probability $1 - \exp(-\alpha^2)$, p appears in at most $O(s \cdot \max(L, \alpha^2))$ of the $D_{i,j}$. The lemma then follows from a union bound over all vertices in B. □

Summarizing the construction and analysis so far

For convenience, we reindex pairs $(C_{i,j}, D_{i,j})$ with $[\ell]$, $\ell = |T^*|$, lexicographically, and denote this bipartite system in H as $((C_i, D_i) : i \in [\ell])$. Assume that θ and R satisfy (4.4). By combining all lemmas so far, with probability $1 - \exp(-\Omega(\alpha^2))$, $((C_i, D_i) : i \in [\ell])$ satisfies:

(a) $\ell = \Omega(\rho^6\gamma R/\alpha^5)$, and the C_i are pairwise disjoint (so it is a bipartite system).

(b) For all $i \in [\ell]$, $h(1/2 - \rho^4/\alpha^4) \leq |C_i| \leq h$ and $|E(C_i, D_i)| \geq \Omega(\rho^4 hk/\alpha^3)$.

(c) For all $i \in [\ell]$ and $p \in D_i$, $|E(p, C_i)| < 2d$.

(d) Every point $p \in D$ appears in at most $O(s \cdot \max(L, \alpha^2))$ of the D_i, $i \in [\ell]$.

So, items 1, 2, and 4 of Definition 4.2.7 remain incomplete.

Item 1 is easy to fix. Let h^* denote the positive integer that maximizes the total number of C_i such that $h^*/2 \leq |C_i| \leq h^*$. By (b) $h^* = \Theta(h)$, and the number of C_i with $h^*/2 \leq |C_i| \leq h^*$ is $\Theta(\ell)$. We only keep such sets C_i in the system. For convenience we use the same ℓ to denote the number of C_i that remains. Then we get a bipartite system $((C_i, D_i) : i \in [\ell])$ in H that satisfies (a), (c), (d) and

(b′) For all $i \in [\ell]$, we have $h^*/2 \leq |C_i| \leq h^*$ and $|E(C_i, D_i)| \geq \Omega(\rho^4 hk/\alpha^3)$.

Cleaning up

Finally, we clean up $((C_i, D_i) : i \in [\ell])$ to obtain a new bipartite system that meets all conditions of Definition 4.2.7, with appropriate parameters. The process consists of three steps (i), (ii), (iii) of further regularizing $((C_i, D_i) : i \in [\ell])$.

First we say the bipartite system $((C_i, D_i) : i \in [\ell])$ is of *degree* g if g is a positive power of 2 that maximizes the following sum: $\sum_{i,p} |C_i \cap N(p)|$, over all $i \in [\ell]$ and $p \in D_i$ with $g \leq |C_i \cap N(p)| < 2g$.

By Property (c) of $((C_i, D_i) : i \in [\ell])$ above, its degree satisfies $g \leq 2d$. Step (i) then removes from each D_i all vertices except those satisfying $g \leq |C_i \cap N(p)| < 2g$. Let D_i' denote the set of vertices left in D_i for each $i \in [\ell]$. Then after Step (i), item 2 of Definition 4.2.7 is now satisfied. While the second part of (b') no longer holds for every $|E(C_i, D_i')|$, we have the following lemma concerning $\sum_i |E(C_i, D_i')|$.

Lemma 4.2.23. *After Step* (i)*, the bipartite system* $((C_i, D_i') : i \in [\ell])$ *satisfies*

$$\sum_{i \in [\ell]} \left|E(C_i, D_i')\right| = \Omega\left(\frac{\rho^4 \ell hk}{\alpha^4}\right).$$

For each $i \in [\ell]$ *and* $p \in D_i'$*, we have* $g \leq |E(p, C_i)| < 2g$*. Moreover, we have*

$$\sum_{i \in [\ell]} |D_i'| = \Omega\left(\frac{\rho^4 \ell hk}{\alpha^4 g}\right).$$

We note that the bipartite system $((C_i, D_i') : i \in [\ell])$ at this moment satisfies (a), items 1 and 2 of Definition 4.2.7 (with m set to h^*, h set to $2g$), Lemma 4.2.23, and (d).

Next, we define a bipartite graph Q, and use it to further clean up the bipartite system $((C_i, D'_i) : i \in [\ell])$ to meet items 3 and 4 of Definition 4.2.7. Each vertex i on the left side of Q corresponds to a set C_i; the right side of Q is exactly $D' = \bigcup_i D'_i$. There is an edge between i and p if $p \in D'_i$. Note that every edge corresponds to roughly g edges in the current system. Denote the number of edges in Q by $N = \sum_{i\in[\ell]} |D'_i|$.

In Step (ii) of the cleaning up we find a positive power r of 2 that maximizes the total degree of vertices on the right side of Q with degree between r and $2r-1$. We remove all vertices on the right side (and their incident edges as well) from Q except those of degree between r and $2r-1$.

Let Q^* denote the new bipartite graph after Step (ii), $D^* \subseteq D$ denote the set of vertices on the right side of Q^*. Let N^* denote the number of edges left in Q^*. From our choice of r, $N^* \geq N/\alpha$. All vertices of Q^* in D^* now have degree between r and $2r-1$. Property (d) implies that $r = O(s \cdot \max(L, \alpha^2))$.

Let LDEG and RDEG denote the average degree of the left and right side of Q^*: LDEG $= N^*/\ell$ and $r \leq$ RDEG $= N^*/|D^*| < 2r$. In Step (iii), we run a deterministic procedure on Q^*:

1. Remove vertices (and incident edges) on the left side of degree $<$ LDEG$/4$.
2. Remove vertices (and incident edges) on the right side of degree $<$ RDEG$/4$.
3. Go back to line 1 if there are still vertices on the left side of degree $<$ LDEG$/4$ or vertices on the right side of degree $<$ RDEG$/4$.

The procedure clearly terminates in polynomial time. Upon termination, each vertex on the left side has degree $\geq$ LDEG$/4$, and each vertex on the right side has degree $\geq$ RDEG$/4$ (but still $< 2r$). Let $\widetilde{Q}$ denote the new graph after Step (iii). Let I denote the set of $i \in [\ell]$ that remains in $\widetilde{Q}$ on the left side, and $\widetilde{D}$ denote the set of vertices that remain on the right side. We will use I and $\widetilde{D}$ to finally obtain a bipartite system that satisfies Definition 4.2.7. Before that we prove the following lemma, showing that the graph still has a lot of edges after Step (iii).

Lemma 4.2.24. *After* Step (iii)*, the number of edges left in* $\widetilde{Q}$ *is at least* $N^*/2$.

Proof. Even if Line 1 of (iii) removes all vertices on the left side of Q^* it can only remove at most $\ell \cdot (\textsc{Ldeg}/4) = N^*/4$ edges. Similarly, even if Line 2 of (iii) removes all vertices on the right side of Q^*, it can only remove at most $N^*/4$ edges in Q^*. The lemma then follows. □

Using $\widetilde{Q}, I$ and $\widetilde{D}$, we define a new bipartite system $((C_i, \widetilde{D}_i) : i \in I)$, where

$$\widetilde{D}_i = D_i \cap \widetilde{D}, \quad \text{for each } i \in I.$$

We show that it is indeed a highly regular bipartite system, and satisfies all items of Definition 4.2.7 with appropriate parameters.

Lemma 4.2.25. *$((C_i, \widetilde{D}_i) : i \in I)$ is a $(\gamma', m', h', t', \rho')$-bipartite system with*

$$\gamma' = \Omega\left(\frac{\rho^{10}}{\alpha^{10}} \cdot \gamma R\right), \quad m' = h^* = \Theta(h), \quad h' = 2g, \quad t' = 2r, \quad \text{and} \quad \rho' = \Omega\left(\frac{\rho^4}{\alpha^5}\right).$$

Proof. First, the C_i are pairwise disjoint and satisfy $h^*/2 \le |C_i| \le h^*$ by (b′).

Second, for all i and $p \in \widetilde{D}_i$, $|E(p, C_i)|$ is between g and $2g - 1$.

For each $i \in I$, the number of edges between $C_i, \widetilde{D}_i$ is $\Omega(\textsc{Ldeg} \cdot g) = \Omega(\rho^4 hk/\alpha^5)$.

Also each $p \in \widetilde{D}$ appears in less than $2r$ but at least $r/4$ of the $\widetilde{D}_i$.

Since the degree of a vertex is k, the degree of $i \in I$ in $\widetilde{Q}$ is $O(hk/g)$. Thus,

$$|I| = \Omega\left(\frac{N^* g}{hk}\right) = \Omega\left(\frac{\rho^4 \ell}{\alpha^5}\right) = \Omega\left(\frac{\rho^{10}\gamma R}{\alpha^{10}}\right).$$

We summarize all properties of the partition operation in the next theorem:

Theorem 4.2.26. *Let $((A_i, B_i) : i \in [\gamma])$ denote a (γ, m, h, t, ρ)-bipartite system in H of type d and strength s. Then d and s satisfy either $d = s = 1$ or*

$$ds = O\big(\alpha^3 ht \cdot \max(\lambda, \mu)/(\rho^4 k)\big).$$

Let θ be a positive integer with $\theta, R = mk\theta/(hn)$ satisfying (4.4) and $L = thk\theta/(dsn)$.

Given a random sequence $y_1, \ldots, y_\theta$ of θ vertices sampled from V, with probability $1 - \exp(-\Omega(\alpha^2))$, the procedure described in this subsection constructs from $((A_i, B_i))$ a $(\gamma', m', h', t', \rho')$-bipartite system $((C_i, \widetilde{D}_i) : i \in [\gamma'])$, where the parameters satisfy

$$\gamma' = \Omega\left(\frac{\rho^{10}}{\alpha^{10}} \cdot \gamma R\right), \ m' = \Theta(h), \quad h' = O(d), \quad t' = O\big(s \cdot \max(L, \alpha^2)\big), \ \rho' = \Omega\left(\frac{\rho^4}{\alpha^5}\right).$$

4.2.3 Interaction

Let $((A_i, B_i) : i \in [\gamma])$ be a (γ, m, h, t, ρ)-bipartite system, and let $\mathbf{X} = (x_1, \ldots, x_\theta)$ denote a tuple of θ vertices drawn from G. Given any $i \in [\gamma]$ and $j \in [\theta]$, $\mathsf{Interact}\,(G, (A_i, B_i), \mathbf{X}, (i, j))$ returns 1 if $x_j \in B_j$, and 0 otherwise.

We prove following lemma for interaction operation.

Lemma 4.2.27. *Let G be a SR graph. Let $((A_i, B_i) : i \in [\gamma])$ be a (γ, m, h, t, ρ)-bipartite system, and $((C_i, D_i) : i \in [\gamma'])$ be a $(\gamma', m', h', t', \rho')$-bipartite system. If*

$$|(\cup_i A_i) \cap (\cup_i C_i)| = \widetilde{O}\left(\frac{\gamma m}{(n/k)^{1/4}}\right), \quad \frac{\gamma m}{ht} = \widetilde{\Omega}\left(\theta\right), \quad m = \widetilde{O}\left(\frac{(n/k)^{1/2}}{\log^a n}\right), \quad \text{and} \quad h = O(1) \tag{4.5}$$

Then at least one of following conditions hold

1. *$\gamma \neq \gamma'$.*
2. *There is an $i \in [\gamma]$ such that $|A_i| \neq |C_i|$.*
3. *Let a be a sufficiently large constant. If $\theta = \lceil (n/k)^{1/2} \cdot \log^a n \rceil$ vertices $\mathbf{X} = (x_1, \ldots, x_\theta)$ are sampled uniformly at random from V, then with probability at least $1 - \exp(-\Omega(\alpha^2))$, there exists a pair $i \in [\gamma]$ and $j \in [\theta]$ such that*

$$\mathsf{Interact}\,(G, ((A_i, B_i)), \mathbf{X}, (i, j)) \neq \mathsf{Interact}\,(G, ((C_i, D_i)), \mathbf{X}, (i, j)) \tag{4.6}$$

Proof. First of all, it must be the case that $\gamma' = \gamma$ and $|A_i| = |C_i|$ for all i; otherwise we are done. Suppose this is indeed the case. Let $A = \cup_i A_i$ and $C = \cup_i C_i$. Using (4.5), we have $|A \cap C| = \widetilde{O}(\gamma m/(n/k)^{1/4})$. So at least $\gamma/2$ many $i \in [\gamma]$ satisfy

$$|A_i \cap C_i| \leq \widetilde{O}\big(m/(n/k)^{1/4}\big). \tag{4.7}$$

Let $I \subseteq [\gamma]$ denote the set of such indexes i. For each $i \in I$ we show below that $|B_i \cap D_i|$ is small. First, $|B_i| \geq \rho mk/h = \Omega(\rho mk)$ as $h = O(1)$. Using (4.7),

$$\big|N(A_i) \cap N(C_i)\big| \leq |A_i \cap C_i|k + |A_i||C_i| \max(\lambda, \mu) = \widetilde{O}\left(\frac{mk}{(n/k)^{1/4}} + m^2 \max(\lambda, \mu)\right)$$

which is clearly an upper bound for $|B_i \cap D_i|$ since $B_i \subseteq N(A_i)$ and $D_i \subseteq N(C_i)$. It follows from Eq. (4.5) and Theorem 4.1.4 that the right hand side is $\ll \rho mk$, if we choose a sufficiently large a. As a result, $|B_i \setminus D_i| = \Omega\left(\rho mk\right)$ for every $i \in I$.

Finally, using the definition of (γ, m, h, t, ρ)-bipartite systems we have

$$\left| \bigcup_{i \in I} (B_i \setminus D_i) \right| \geq \Omega\left(\frac{|I| \cdot \rho m k}{ht}\right) = \Omega\left(\frac{\rho \gamma m k}{ht}\right) = \widetilde{\Omega}\left(k\theta\right),$$

using (4.5). As $k\theta \cdot (\theta/n) = \log^{2a} n$, by choosing a sufficiently large constant a we can guarantee with probability $1 - \exp(-\Omega(\alpha^2))$, there are j and i such that $x_j \in B_i \setminus D_i$. □

4.2.4 A canonical pairwise distinguisher for SR graphs

We are now ready to present the **Cert** function, based on the partition operation and interaction operation. Details of **Cert** function are presented in Figure 4.1. It follows from the description of the partition operation that **Cert** (for parameters c and θ specified in Eq. (4.11)) is a polynomial-time computable operator.

For the rest of this section, we fix $\varepsilon > 0$ to be any positive constant and use $\mathcal{K}$ to denote the set of non-trivial SR graphs satisfying $k = \Omega(n^{2/3})$ and $k = O(n^{1-\varepsilon})$.

Lemma 4.2.28. *Let G be a graph in $\mathcal{K}$, and*

$$\theta = \left\lceil (n/k)^{1/2} \cdot \log^a n \right\rceil \quad \text{and} \quad c = \lceil 4/\varepsilon \rceil, \tag{4.11}$$

where a denotes a sufficiently large constant to be specified later. There exists a map $f : [(c+1)\theta] \to V$ such that

$$\mathbf{Cert}\,(G, f, x, y) \neq \text{nil}, \quad \textit{for any two distinct vertices } x, y \in V.$$

We prove Lemma 4.2.28 in the rest of this section. For clarity of the argument we state additional lemmas within the proof.

Proof of Lemma 4.2.28. Fix $u \neq v \in V$. We show below that if mapping f is sampled randomly, then $\mathbf{Cert}(G, f, u, v)$ is not nil with high probability. Lemma 4.2.28 then follows by a union bound.

To this end, we first have the following direct corollary of Theorem 4.2.26.

Corollary 4.2.29. *The probability that* $\mathbf{Cert}(G, f, u, v)$ *halts because the construction of a bipartite system fails in Step 2 is at most* $\exp(-\Omega(\alpha^2))$.

Input: $G = (V, E)$ in $\mathcal{K}$, $u, v \in V$, $f : [(c+1)\theta] \to V$

0. Set $y_{i,j}$ to be the vertex $f((i-1)\theta + j)$; set x_j to be the vertex $f(c\theta + j)$.
1. Let $\gamma_0 = 1, m_0 = k, h_0 = \mu, t_0 = 1, \rho_0 = 1 - o(1)$. $\mathcal{S}_0 = ((N(u), V \setminus N^+(u)))$ is a bipartite system in H_u, with the above parameters. Let d_0 denote the type and s_0 denote the strength of $\mathcal{S}_0$. Similarly construct $\mathcal{S}'_0$ for vertex v, and let $\gamma'_0 = 1$, $m'_0 = k$, $h'_0 = \mu$, $t'_0 = 1$, $\rho'_0 = 1 - o(1)$. Set $r = 0$.
2. If $r < c$ and θ satisfies:
$$\theta = O\left(\frac{\rho_r^{10} n}{\alpha^{10} k}\right) \quad \text{and} \quad R_r = \frac{m_r k \theta}{h_r n} \geq \frac{\alpha^8}{\rho_r^6}, \tag{4.8}$$
we use the partition operation described in the previous subsection to build a new $(\gamma_{r+1}, m_{r+1}, h_{r+1}, t_{r+1}, \rho_{r+1})$-bipartite system $\mathcal{S}_{r+1}$ in H_u, using $\mathcal{S}_r$ and θ samples $y_{r+1,1}, \ldots, y_{r+1,\theta}$. Let d_r, s_r denote the type and strength of the old bipartite system $\mathcal{S}_r$. Similarly construct $(\gamma'_{r+1}, m'_{r+1}, h'_{r+1}, t'_{r+1}, \rho'_{r+1})$-bipartite system $\mathcal{S}'_r$ for vertex v.
If parameters of $\mathcal{S}_r$ satisfy
$$d_r = s_r = 1 \quad \text{and} \quad \frac{t_r h_r k \theta}{n} \leq \alpha^2 \tag{4.9}$$
or parameters of $\mathcal{S}'_r$ satisfy
$$d'_r = s'_r = 1 \quad \text{and} \quad \frac{t'_r h'_r k \theta}{n} \leq \alpha^2 \tag{4.10}$$
increment r and go to Step 3; otherwise, increment r and go back to Step 2.
If at the beginning parameters of $\mathcal{S}_r$(or $\mathcal{S}'_r$) violate (4.8), go to Step 3. If $r = c$ (running out of samples, which we will show never happens) or the partition operation fails, halt and return nil.
3. Let $\mathcal{S}_r = ((A_i, B_i) : i \in [\gamma_r])$ and $\mathcal{S}'_r = ((C_i, D_i) : i \in [\gamma'_r])$. Then
   (a) Return 0 if $\gamma_r \neq \gamma'_r$; Return i if $|A_i| \neq |C_i|$;
   (b) a pair (i, j) if there exists $i \in [\gamma_r]$ and $j \in [\theta]$ such that
   $$\mathsf{Interact}\,(G, (A_i, B_i), (x_1, \ldots, x_\theta), (i, j)) \neq \mathsf{Interact}\,(G, (C_i, D_i), (x_1, \ldots, x_\theta), (i, j))\,;$$
   (c) Return nil otherwise.

Figure 4.1: The algorithm **Cert** for $\exp(\widetilde{O}(\sqrt{n/k}))$ bound of SR graphs

Assume that $\mathbf{Cert}(G, f, u, v)$ successively constructed two sequences of $\ell + 1$ bipartite structures in Step 2: $\mathcal{S}_0, \ldots, \mathcal{S}_\ell$ and $\mathcal{S}'_0, \ldots, \mathcal{S}'_\ell$, for some $\ell \le c$, and then either halts because $\ell = c$ (running out of samples) or moves to Step 3 because (4.9) is satisfied by parameters of $\mathcal{S}_{\ell-1}$(or $\mathcal{S}'_{\ell-1}$) or (4.8) is violated by parameters of $\mathcal{S}_\ell$(or $\mathcal{S}'_{\ell-1}$).

We let $(\gamma_i, m_i, h_i, t_i, \rho_i)$ denote the parameters of $\mathcal{S}_i$, with type d_i and strength s_i, and $(\gamma'_i, m'_i, h'_i, t'_i, \rho'_i)$ denote the parameters of $\mathcal{S}'_i$, with type d'_i and strength s'_i. Let $\mathcal{S}_i = ((A_{i,j}, B_{i,j}))$, $A_i = \bigcup_j A_{i,j}$, and $B_i = \bigcup_j B_{i,j}$, and $\mathcal{S}'_i = ((C_{i,j}, D_{i,j}))$, $C_i = \bigcup_j C_{i,j}$, and $D_i = \bigcup_j D_{i,j}$, Let $W = N(u) \cap N(v)$ and $W_i = A_i \cap W$ (so $W_0 = W$).

We prove the following lemma about these parameters for $\mathcal{S}_i$. The same bounds applies to $\mathcal{S}'_i$.

Lemma 4.2.30. *$\mathcal{S}_0, \ldots, \mathcal{S}_\ell$ satisfy the following conditions: for each $i \in [0 : \ell - 1]$,*

$$\gamma_{i+1} = \Omega\left(\frac{\rho_i^{10}}{\alpha^{10}} \cdot \gamma_i \cdot \frac{m_i k\theta}{h_i n}\right), \quad m_{i+1} = \Theta(h_i), \quad h_{i+1} = O(d_i). \tag{4.12}$$

We also have

$$t_{i+1} = O\left(\max\left(\alpha^2 s_i, \frac{h_i t_i k\theta}{d_i n}\right)\right), \quad \rho_{i+1} = \Omega\left(\frac{\rho_i^4}{\alpha^5}\right), \quad |W_{i+1}| \le |W_i| \cdot \frac{\theta k}{n} + \alpha\sqrt{\theta} h_i t_i. \tag{4.13}$$

Proof. Immediate from Theorem 4.2.26 and Lemma 4.2.21. □

Next we show that $\ell = c$ (running out of samples) never happens.

Lemma 4.2.31. $\mathbf{Cert}(G, f, u, v)$ *never halts due to $\ell = c$ for $\mathcal{S}_0, \ldots, \mathcal{S}_\ell$ in Step 2.*

Proof. By induction and (4.12), ρ_i is $\Omega(1/\mathrm{polylog}(n))$ for all i (as c is a constant).

Next by (4.12), $m_{i+1} = \Theta(h_i)$ for all $i < \ell$. Using Lemma 4.2.9, either $h_i = O(1)$ or $h_i = O(m_i \cdot \max(\lambda, \mu)/(\rho_i k))$. So either $m_{i+1} = O(1)$ or it drops by a factor of

$$O\left(\frac{\max(\lambda, \mu)}{\rho_r k}\right) = \widetilde{O}\left((k/n)^{1/2}\right) = \widetilde{O}\big(n^{-\varepsilon/2}\big),$$

from m_i, using Theorem 4.1.4 and Lemma 4.1.3. Since $m_0 = k$, within the $c = \lceil 4/\varepsilon \rceil$ rounds there must be a round in which $m_i = O(1)$. Let j be the first such round, $R_j = O(k\theta/(h_j n)) \ll 1$ and $\mathbf{Cert}(G, f, u, v)$ moves to Step 3, a contradiction. □

Now we know that **Cert**(G, f, u, v) reaches Step 3 with high probability. We analyze carefully final parameters of $\mathcal{S}_\ell$. As mentioned earlier, $\rho_i = \Omega(1/\text{polylog}(n))$ for all i. Using $\gamma_0 = 1$, $m_0 = k$, $m_{i+1} = \Theta(h_i)$ and (4.12), we have by induction for any $i \geq 0$:

$$\gamma_i m_i = \Omega\left(\frac{1}{\text{polylog}(n)} \cdot \left(\frac{\theta k}{n}\right)^i \cdot k\right). \tag{4.14}$$

where the exponent of $\log n$ in polylog(n) is bounded, as there are at most c rounds. From now on we use $\widetilde{O}$ and $\widetilde{\Omega}$ to suppress polylog factors (by the expression $f = \widetilde{\Omega}(g)$ we mean $g = \widetilde{O}(f)$).

We study the first case: **Cert** reaches Step 3 because (4.8) is violated. Since $\rho_i = \Omega(1/\text{polylog}(n))$ and $k = O(n^{1-\varepsilon})$, the first condition on θ in (4.8) always holds. So it must be the case that the second condition in (4.8) is violated. Our goal is then to prove the following set of bounds on parameters of $\mathcal{S}_\ell$:

$$\frac{\gamma_\ell m_\ell}{h_\ell t_\ell} = \widetilde{\Omega}\left(k/\mu\right), \quad m_\ell = \widetilde{O}\left(\frac{(n/k)^{1/2}}{\log^a n}\right), \quad \text{and} \quad h_\ell = O(1). \tag{4.15}$$

To prove the first one we would like to show for i from 0 to $\ell - 1$, one never needs to invoke the max in the upper bound of t_{i+1} in (4.13) but always have

$$t_{i+1} = \widetilde{O}\left(\frac{h_i t_i k\theta}{d_i n}\right). \tag{4.16}$$

If this is indeed the case then the first equation of (4.15) follows easily from bounds on $\gamma_{i+1}, m_{i+1}, h_{i+1}$ and r_{i+1} in (4.12) and (4.13) and an induction on i.

To see (4.16), we focus on the case when $d_i s_i > 1$ (as (4.16) is trivial if $d_i s_i = 1$ but parameters of $\mathcal{S}_i$ violates the second bound of (4.9)) and Lemma 4.2.15 to get

$$\frac{t_i h_i k\theta}{d_i s_i n} = \frac{\theta k}{n} \cdot t_i h_i \cdot \widetilde{\Omega}\left(\frac{k}{h_i t_i \cdot \max(\lambda, \mu)}\right) = \widetilde{\Omega}\big(\log^a n\big),$$

where we used $\lambda = O(\sqrt{k\mu})$ from Theorem 4.1.4 since $k = \Omega(n^{2/3})$. Note that the exponent of $\log n$ in the hidden polylog factor is a constant that depends on c only but is independent of our choice of a (so later we can pick a sufficiently large a to suppress it if desired).

To prove the other two bounds of (4.15), we combine the violation of the second condition in (4.8) by $\mathcal{S}_\ell$ with Lemma 4.2.9. If h_ℓ is not $O(1)$, we have

$$m_\ell \leq \frac{\alpha^8}{\rho_\ell^6} \cdot \frac{h_\ell n}{\theta k} = \text{polylog}(n) \cdot O\left(\frac{m_\ell \cdot \max(\lambda, \mu))}{\rho_\ell k}\right) \cdot \frac{(n/k)^{1/2}}{\log^a n}.$$

If we choose a to be sufficiently large, this cannot happen. So $h_\ell = O(1)$ and

$$m_\ell \leq \frac{\alpha^8}{\rho_\ell^6} \cdot \frac{h_\ell n}{\theta k} = \widetilde{O}\left(\frac{(n/k)^{1/2}}{\log^a n}\right). \tag{4.17}$$

Next we work on the other case when parameters of $\mathcal{S}_{\ell-1}$ satisfy (4.9). Our goal is the following bounds on $\mathcal{S}_\ell$ (the first one is weaker but the other two are the same):

$$\frac{\gamma_\ell m_\ell}{h_\ell t_\ell} = \widetilde{\Omega}(\theta), \quad m_\ell = \widetilde{O}\left(\frac{(n/k)^{1/2}}{\log^a n}\right), \quad \text{and} \quad h_\ell = O(1) \tag{4.18}$$

As $\mathcal{S}_{\ell-1}$ is the first bipartite system in the sequence that satisfies (4.9), an argument similar to that used in proving the first bound of Eq. (4.15) gives us

$$\frac{\gamma_{\ell-1} m_{\ell-1}}{h_{\ell-1} t_{\ell-1}} = \widetilde{\Omega}(k/\mu)$$

and trivially $\gamma_{\ell-1} m_{\ell-1} = \widetilde{\Omega}(k/\mu)$. From (4.12), (4.13) and (4.9) we have $h_\ell = O(1)$ and $t_\ell = O(\alpha^2)$ as well as $m_\ell = \Theta(h_{\ell-1}) = O(\alpha^2 n/(k\theta))$. Combining (4.12) and the bound on $\gamma_{\ell-1} m_{\ell-1}$ we get all three bounds claimed in Eq. (4.18). Since the first bound in (4.18) is weaker while the other two bounds are the same in (4.15) and (4.18) we will use the latter on parameters of $\mathcal{S}_\ell$.

For $\mathcal{S}'_\ell$, we have same bounds. Hence, Lemma 4.2.28 then follows from Lemma 4.2.27 by a union bound. □

4.3 Automorphism of SR graphs

In last section, we showed a new algorithm for SR graph isomorphism with running time $\exp(\widetilde{O}(n^{1/5}))$. However, this result does not imply the same upper bound for the order of automorphism groups for non-trivial non-graphical SR graphs, because (a) of Theorem 4.2.4 relies on group theory method, which does not yield bounds for the order of automorphism group.

In this section, we present the following bound for non-trivial non-graphical SR graphs, improving previous $\exp(\widetilde{O}(n^{1/3}))$ upper bound by Spielman [Spielman, 1996].

Theorem 4.3.1. *Every non-trivial non-graphical SR graph G has $|\operatorname{Aut}(G)| = \exp(\widetilde{O}(n^{9/37}))$.*

4.3.1 Latin square graphs and Steiner graphs

Proposition 4.3.2. *Let G be a strongly regular graph with parameters (n, k, λ, μ). Let s denote the smallest eigenvalue as in* Theorem 4.1.2. *Assume n is bounded below by a sufficiently large constant. If G does not satisfy the claw bound and $k < n/\log n$, then one of following two conditions holds:*

1. *either G is a Steiner graph derived from a Steiner 2-design that satisfies $\sqrt{n} - 2 > (-s-1)^2$;*
2. *or G is a Latin square graph derived from an s-net with $n > (-s-1)^4$.*

Proof. Let G be a strongly regular graph which does not satisfy Neumaier's claw bound. As observed in Spielman [Spielman, 1996], Neumaier's characterization (Theorem 4.1.2) states that G is either a Steiner graph or a Latin square graph, depending on whether $\mu = s^2$ or $\mu = s(s+1)$.

If G is a Steiner graph obtained from a Steiner 2-design, then we have $r = \lambda - \mu - s = \lambda - s^2 - s$. As $\lambda = h - 2 + (-s-1)^2$ where h denotes the number of lines passing through each point in the Steiner 2-design (e.g., see the proof of Proposition 10 in [Spielman, 1996]), we have

$$h - 2 + (-s-1)^2 - s^2 - s = r > 2(-s-1)(s^2 + 1 + s)$$

where the inequality follows from the violation of claw bound. As $\sqrt{n} \geq h/\sqrt{2} > h/2$ from (3.2), we get

$$\sqrt{n} - 2 > \frac{h}{2} - 2 > \frac{(-s-1)(2s^2+2s+3)+2}{2} - 2 > (-s-1)^2$$

where the last inequality always holds for any negative integer $s < -1$.

If G is a Latin square graph obtained from an s-net, then the violation of claw bound implies that

$$k > r > \frac{s(s+1)(\mu+1)}{2} - s - 1 \geq (-s-1)^4/2$$

Since $n > k \cdot \log n$, we have $n > (-s-1)^4$ when n is sufficiently large. The lemma follows. □

For the case of *s-net*, Miller proved following theorem in [Miller, 1978].

Theorem 4.3.3. *Every s-net with n points completely splits by $\log n$ points under naive vertex refinement.*

4.3.2 A canonical pairwise distinguisher for SR graphs with the claw bound

In this section, we present an algorithm for general strongly regular graphs that satisfy *the claw bound* and

$$k \leq n^{7/13} / \log n. \tag{4.19}$$

In the description of the algorithm and its analysis, we use α, β and γ to denote the following integers:

$$\alpha = \left\lceil \frac{n^{7/3}}{k^{13/3} \cdot \log n} \right\rceil, \quad \beta = \left\lceil \frac{\alpha}{\log n} \right\rceil, \quad \text{and} \quad \gamma = \left\lceil \frac{k^{17/3}}{n^{8/3}} \cdot \log^5 n \right\rceil$$

From (4.19) and $k \geq \sqrt{n-1}$, we have

$$\Omega\big(\log^{10/3} n\big) \leq \alpha \leq O\big(n^{1/6}/\log n\big), \quad \beta = \Omega\big(\log^{7/3} n\big) \quad \text{and} \quad \gamma = \Omega\big(n^{1/6} \cdot \log^5 n\big)$$

We also use the following notation. Given a vertex u in G, let $N(u)$ denote the set of neighbors of u and $N^+(u)$ denote $N(u) \cup \{u\}$. Thus, $|N(u)| = k$ and $|N^+(u)| = k+1$. Then, for a set of vertices A in G, $N(A)$ and $N^+(A)$ will denote $\cup_{u \in A} N(u)$ and $\cup_{u \in A} N^+(u)$, respectively.

Our algorithm in this section relies on two deterministic (polynomial-time algorithms) also called **Test** and **Cert**. Details of these two algorithms can be found in Figure 4.2.

From the description of **Test** and **Cert** in Figure 4.2, **Test** and **Cert** form a canonical pairwise distinguisher for the graph.

In the rest of this section, we will prove the following main technical theorem.

Theorem 4.3.4. *Let $G = (V, E)$ be a strongly regular graph with parameters (n, k, μ, λ) satisfying the claw bound and (4.19). Let (u, v) be a pair of distinct vertices in G. If a map f from $[\alpha + \beta + \gamma]$ to V is sampled uniformly at random, then $\mathbf{Cert}(G, f, u, v) \neq nil$ with probability at least $1 - \exp(-\Omega(\log^2 n))$.*

By requiring

$$k \leq n^{19/37} \ll n^{7/13} / \log n$$

we have

$$\exp\Big(\tilde{O}(\alpha + \beta + \gamma)\Big) = \exp\Big(\tilde{O}(n^{1/6} + n^{1/6} + n^{9/37})\Big) = \exp\Big(\tilde{O}(n^{9/37})\Big)$$

Algorithm Cert (G, f, u, v)

Input: $G = (V, E)$ is a strongly regular graph with parameter (n, k, μ, λ) satisfying the claw bound and (4.19); u and v are two distinct vertices in G; f is a map from $[\alpha + \beta + \gamma]$ to V.

1. Break f into three maps $f_1 : [\alpha] \to V$, $f_2 : [\beta] \to V$ and $f_3 : [\gamma] \to V$ such that

$$f_1(i) = f(i), \quad f_2(j) = f(\alpha + j) \quad \text{and} \quad f_3(\ell) = f(\alpha + \beta + \ell)$$

2. [*Expansion*]: For each $i \in [\alpha]$, construct $A_{u,i}$ as follows: If $f_1(i) \in N^+(u)$, then set $A_{u,i} = \emptyset$; otherwise $A_{u,i}$ is the set of μ common neighbors of u and $f_1(i)$. For each $i \in [\alpha]$ and $j \in [\beta]$, construct $B_{u,i,j}$ as follows: If $A_{u,i} = \emptyset$ or $f_2(j) \in N^+(A_{u,i})$, set $B_{u,i,j} = \emptyset$; otherwise $B_{u,i,j}$ is the set of all vertices that are common neighbors of $f_2(j)$ and at least one vertex from $A_{u,i}$:

$$B_{u,i,j} = \Big\{ w \in N\big(f_2(j)\big) : \exists\, w' \in A_{u,i} \text{ such that } (w, w') \in E \Big\}$$

   Similarly construct $A_{v,i}$ for each i, and $B_{v,i,j}$ for each pair (i, j), $i \in [\alpha]$ and $j \in [\beta]$.

3. [*Interaction*]: If there exists a triple (i, j, ℓ), where $i \in [\alpha]$, $j \in [\beta]$ and $\ell \in [\gamma]$, such that

$$f_3(\ell) \in N(B_{u,i,j}) \quad \text{but} \quad f_3(\ell) \notin N(B_{v,i,j})$$

   return (i, j, ℓ); otherwise, return nil.

Algorithm Test $(G, f, (i, j, \ell), u)$

Input: $G = (V, E)$ is a strongly regular graph with parameter (n, k, μ, λ) satisfying the claw bound and (4.19); u is a vertex in G; f is a map from $[\alpha + \beta + \gamma]$ to V; and (i, j, ℓ) is a triple of integers that satisfy $i \in [\alpha]$, $j \in [\beta]$, and $\ell \in [\gamma]$.

1. Return 1 if $f_3(\ell) \in N(B_{u,i,j})$; and return 0 otherwise.

Figure 4.2: The two algorithms **Test** and **Cert** for $\exp(\widetilde{O}(n^{9/37}))$ bound of SR graphs

Before proving Theorem 4.3.4, we introduce a definition and establish a few lemmas that will be useful for the proof. In the rest of the section, we always assume G satisfies the claw bound as well as (4.19).

Definition 4.3.5 (A Good Start)**.** *Let (u, v) be a pair of distinct vertices in G, and f_1 be a map from $[\alpha]$ to V. We say f_1 is* good with respect to (u, v) *if there exists a set $I \subseteq [\alpha]$ of size $|I| \geq \alpha/2$ such that*

1. *$A_{u,i} \neq \emptyset$ and $A_{v,i} \neq \emptyset$, for all $i \in I$;*
2. *$A_{u,i} \cap A_{u,j} = \emptyset$ and there is no edge between $A_{u,i}$ and $A_{u,j}$, for all $i, j : i \neq j \in I$;*
3. *$A_{v,i} \cap A_{v,j} = \emptyset$ and there is no edge between $A_{v,i}$ and $A_{v,j}$, for all $i, j : i \neq j \in I$; and*
4. *$A_{u,i} \cap A_{v,i} = \emptyset$, for all $i \in I$.*

We first show that f_1 is *good* with high probability, if it is sampled uniformly at random.

Lemma 4.3.6. *Let (u, v) be a pair of distinct vertices. If a map f_1 from $[\alpha]$ to V is sampled uniformly at random, then f is good with probability $1 - \exp(-\Omega(\alpha))$, and one can compute in polynomial time a set $I \subseteq [\alpha]$ that satisfies all the conditions of* Definition 4.3.5.

Proof. We construct I as follows. Start with $I = \emptyset$, and sample $f(1), f(2), \ldots, f(\alpha)$ one by one. For each $i \geq 1$, assume $f(1), f(2), \ldots, f(i-1)$ have already been sampled. If the point $f(i)$ we get satisfies all the following four conditions, then we add it to I:

1. $f(i) \notin N^+(u) \cup N^+(v)$;
2. $A_{u,i} \cap A_{v,i} = \emptyset$;
3. $A_{u,i} \cap A_{u,t} = \emptyset$ and there is no edge between $A_{u,i}$ and $A_{u,t}$, for all $t < i$ and $t \in I$;
4. $A_{v,i} \cap A_{v,t} = \emptyset$ and there is no edge between $A_{v,i}$ and $A_{v,t}$, for all $t < i$ and $t \in I$.

By induction it is clear that the set I we get by the end satisfies all the conditions of Definition 4.3.5. We now show that $|I| \geq \alpha/2$ with high probability. Note that the number of vertices in G that, when picked as $f(i)$, violate each of the four conditions can be bounded above respectively by

$$O(k), \quad \max\{\mu,\lambda\}\cdot k, \quad (i-1)\mu k+(i-1)\mu\lambda k \quad \text{and} \quad (i-1)\mu k+(i-1)\mu\lambda k$$

The first two bounds are self-evident. To see the last two bounds, we consider a pair i and t. First note that if $f(i)$ is not connected to $A_{u,t}$, then $A_{u,i}\cap A_{u,t}=\emptyset$. As $|A_{u,t}|=\mu$, $A_{u,t}$ is directly connected to at most μk vertices. Second note that if $f(i)$ is not connected with a common neighbor of u and a member of $A_{u,t}$, then there is no edge between $A_{u,i}$ and $A_{u,t}$. The total number of common neighbors of u and members of $A_{u,t}$ is at most $\mu\lambda$. We therefore obtain the third, and similarly the last, bound.

Since $i\le\alpha$, the probability of i being added to I is at least

$$1-O\left(\frac{k+\max\{\mu,\lambda\}\cdot k+\alpha\mu k+\alpha\mu\lambda k}{n}\right)=1-O\left(\frac{\alpha\mu\lambda k}{n}\right)=1-O\left(\frac{1}{\log n}\right)$$

where $\mu=O(k^2/n)$ and $\lambda=O(k^{4/3}/n^{1/3})$. So the expectation of $|I|$ is $\alpha(1-o(1))$. By Chernoff bound, we have $|I|\ge\alpha/2$ happens with probability at least $1-\exp(-\Omega(\alpha))$. □

From now on we assume that the event described in Lemma 4.3.6 occurs: f_1 has already been sampled and it is good with respect to (u,v); We have obtained a set $I\subseteq[\alpha]$ that satisfies all the conditions of Definition 4.3.5, including $|I|\ge\alpha/2$. Let m denote (note that m could be much smaller than 1):

$$m=\frac{k^{19/3}}{n^{10/3}}\cdot\log n$$

Next we sample f_2, and use $R\subseteq I\times[\beta]$ to denote the following set of pairs (i,j): $(i,j)\in R$ if $i\in I$ and $f_2(j)$ satisfies the following four conditions:

1. $f_2(j)\notin N^+(A_{u,i})\cup N^+(A_{v,i})$;
2. $|B_{u,i,j}\cap B_{v,i,j}|\le m$;
3. There are at most m vertices in $B_{u,i,j}$ that are connected to at least two vertices in $A_{u,i}$;
4. There are at most m vertices in $B_{v,i,j}$ that are connected to at least two vertices in $A_{v,i}$.

We prove that every pair $(i,j)\in R$ has the following property:

Lemma 4.3.7. *If $(i,j) \in R$, then we have $|B_{u,i,j} - B_{v,i,j}| \geq \mu^2/2$ and $|B_{v,i,j} - B_{u,i,j}| \geq \mu^2/2$.*

Proof. Because $f_2(j) \notin N^+(A_{u,i})$, each vertex in $A_{u,i}$ shares μ common neighbors with $f_2(j)$. Hence the total number of pairs (w, w') such that $w \in A_{u,i}$ and $(w, w'), (f_2(j), w') \in E$ is exactly μ^2. Also note that $|B_{u,i,j}|$ is exactly the number of distinct w''s in such pairs. Because $(i,j) \in R$, the number of w''s that appear in at least two such pairs is no more than m. As $|A_{u,i}| = \mu$, each w' can appear in no more than μ different pairs. Together we know the number of w''s that appear in exactly one pair is at least

$$\mu^2 - m\mu = \mu^2\big(1 - o(1)\big)$$

where $m = o(\mu)$ follows from (4.19). Thus, $|B_{u,i,j}| = \mu^2(1-o(1))$ and using $|B_{u,i,j} \cap B_{v,i,j}| \leq m$ we get

$$|B_{u,i,j} - B_{v,i,j}| \geq \mu^2\big(1 - o(1)\big) - m \geq \mu^2/2.$$

Similarly we have $|B_{v,i,j} - B_{u,i,j}| \geq \mu^2/2$. □

We will use the following probabilistic statement about R.

Lemma 4.3.8. *For each $i \in I$, if $f_2(j)$, where $j \in [\beta]$, is sampled uniformly at random, then*

$$\Pr\Big[\,(i,j) \in R\,\Big] = 1 - O\big(1/\log n\big)$$

Proof. Since $i \in I$ and $|A_{u,i}| = |A_{v,i}| = \mu$, we have $|N^+(A_{u,i})| + |N^+(A_{v,i})| = O(\mu k)$. So the probability that $f_2(j)$ violates the first condition is at most $O(\mu k/n) = O(k^3/n^2) = o(1/\log n)$.

To analyze the third condition, we let

$$W = \Big\{ z \in V : \text{there are two distinct } w, w' \in A_{u,i} \text{ such that } (z,w) \text{ and } (z,w') \in E \Big\}$$

As each such w, w' can have at most $\max\{\lambda, \mu\}$ many common neighbors, we have

$$|W| \leq |A_{u,i}|^2 \cdot \max\{\lambda, \mu\} \leq \mu^{7/3} k^{2/3} = O\left(\frac{k^{16/3}}{n^{7/3}}\right)$$

When $f_2(j) \notin N^+(A_{u,i})$, $W \cap N(f_2(j))$ is exactly the set of vertices in $B_{u,i,j}$ that are connected to at least two vertices in $A_{u,i}$. For a randomly sampled vertex $f_2(j)$, the expectation of $|W \cap N(f_2(j))|$ is

$$O\left(\frac{k^{16/3}}{n^{7/3}}\right) \cdot \frac{k}{n} = O\left(\frac{k^{19/3}}{n^{10/3}}\right) = O\left(\frac{m}{\log n}\right) \tag{4.20}$$

It follows that $f_2(j)$ violates the third, and similarly the last condition, with probability $O(1/\log n)$.

Finally, we examine the probability of $|B_{u,i,j} \cap B_{v,i,j}| \geq m$. To this end, we count the number K of vertices that are connected to at least one vertex in $A_{u,i}$ and at least one vertex in $A_{v,i}$. Because $i \in I$, we have $A_{u,i} \cap A_{v,i} = \emptyset$ and thus, K is at most $\mu^2 \cdot \max\{\lambda, \mu\}$. Similarly, for a randomly sampled $f_2(j)$, the expectation of $|B_{u,i,j} \cap B_{v,i,j}|$ can be bounded by (4.20). It follows that the probability of j violating the second condition is also $O(1/\log n)$. The lemma then follows using the union bound. □

For each $j \in [\beta]$, let $I_j \subseteq I$ denote the set of $i \in I$ with $(i, j) \in R$. We need the following definition:

Definition 4.3.9 (A Healthy Second Step)**.** *We say $j \in [\beta]$ is* good *with respect to (u, v) and f_1, if $|I_j| \geq |I|/2 = \Omega(\alpha)$ and for any $i, i' : i \neq i' \in I_j$, $B_{u,i,j} \cap B_{u,i',j} = \emptyset$ and $B_{v,i,j} \cap B_{v,i',j} = \emptyset$.*

Lemma 4.3.10. *If $f_2(j)$, where $j \in [\beta]$, is sampled uniformly at random, then with probability at least $1 - O(1/\log n)$, j is* good *with respect to (u, v) and f_1.*

Proof. We first consider the probability of $B_{u,i,j} \cap B_{u,i',j} \neq \emptyset$, for some $i, i' : i \neq i' \in I$.

When this event happens, there must be a vertex w such that $(w, f_2(j)) \in E$ and w is connected to at least one vertex in $A_{u,i}$ and at least one vertex in $A_{u,i'}$, for some $i \neq i' \in I$. On the other hand, since there is no edge between any two sets $A_{u,i}$ and $A_{u,i'}$, the number of vertices that are connected to more than one sets $\{A_{u,i}\}$, $i \in I$, can be easily bounded by $O((|I|\mu)^2\mu)$. For a randomly sampled vertex $f_2(j)$ the probability of $B_{u,i,j} \cap B_{u,i',j} \neq \emptyset$ for some $i \neq i' \in I$ is at most

$$O\left(|I|^2\mu^3 \cdot \frac{k}{n}\right) = O\left(\frac{n^{2/3}}{k^{5/3} \cdot \log^2 n}\right) = O\left(\frac{1}{n^{1/6} \cdot \log^2 n}\right)$$

So the probability of $B_{u,i,j} \cap B_{u,i',j} = \emptyset$ for all i, i' is $\geq 1 - O(1/n^{1/6})$. The same bound holds for v.

To complete the proof, it suffices to show that $I_j \subseteq I$ is large with high probability. By Lemma 4.3.8 we have the expectation of $|I_j|$ is $|I|(1 - O(1/\log n))$. It then follows that $|I_j| \geq |I|/2$ with probability at least $1 - O(1/\log n)$. The lemma follows using the union bound. □

Lemma 4.3.11 (Ready to Interact)**.** *With probability* $1 - \exp(-\Omega(\beta))$*, there exists a set* $J \subseteq [\beta]$ *of size at least* $\beta/2$ *such that every* $j \in J$ *is* good *with respect to* (u, v) *and map* f_1*; and* $\{f_2(j) : j \in J\}$ *is an independent set in* G *of size* $|J|$*.*

Proof. We construct J by sampling $f_2(1), f_2(2), \ldots, f_2(\beta)$ one by one. Start with $J = \emptyset$. For each $j \in [\beta]$ we add j to J if j is good and $f_2(j) \notin N^+(f_2(1)) \cup \cdots \cup N^+(f_2(j-1))$. By Lemma 4.3.10, the first event happens with probability $1 - O(1/\log n)$. The second event happens with probability at least

$$1 - \frac{(j-1)(k+1)}{n} = 1 - O\left(\frac{\beta k}{n}\right) = 1 - O\left(\frac{1}{n^{1/3} \cdot \log^2 n}\right)$$

Thus, each j is added to J with probability $1 - O(1/\log n)$. The lemma follows by Chernoff bound. □

We now analysis the probability of successful interaction in **Cert**, when both events as described in Lemma 4.3.6 and 4.3.11 occur. Let $I \subseteq [\alpha]$ and $J \subseteq [\beta]$ denote the two sets satisfying the conditions of Lemma 4.3.6 and 4.3.11. For each pair (i, j), where $j \in J$ and $i \in I_j$, we use $C_{u,i,j}$ and $C_{v,i,j}$ to denote

$$C_{u,i,j} = N(B_{u,i,j}) \quad \text{and} \quad C_{v,i,j} = N(B_{v,i,j})$$

Based on Step 3 of **Cert**, if there exists an $\ell \in [\gamma]$ such that $f_3(\gamma) \in C_{u,i,j} - C_{v,i,j}$ then **Cert** will return a triple rather then nil. Letting $C_{u,j} = \cup_{i \in I_j}(C_{u,i,j} - C_{v,i,j})$, and we now prove the following two lemmas to give a lower bound for $|\cup_{j \in J} C_{u,j}|$.

Lemma 4.3.12. *For each* $j \in J$*, we have* $|C_{u,j}| = \Omega(\mu\alpha k)$*.*

Proof. Since $j \in J$, we know that the $B_{u,i,j}$'s are pairwise disjoint for $i \in I_j$. Let

$$W_{(u,v),j} = \cup_{i \in I_j}(B_{u,i,j} - B_{v,i,j})$$

Note that $W_{(u,v),j}$ is a subset of $N(f_2(j))$. By Lemma 4.3.7, it also satisfies

$$|W_{(u,v),j}| = \sum_{i \in I_j} |B_{u,i,j} - B_{v,i,j}| = \Omega(\mu^2 \alpha)$$

To get a lower bound for $|C_{u,j}|$, we consider the following family of $|W_{(u,v),j}|$ sets:

$$\Big\{N(z) - C_{v,i,j} - N^+\big(f_2(j)\big)\Big\}_{i \in I_j \text{ and } z \in B_{u,i,j} - B_{v,i,j}}$$

By definition, each of these sets is a subset of $C_{u,j}$. We first give a lower bound on the size of any set in this family. Consider a vertex $z \in B_{u,i,j} - B_{v,i,j}$ for some $i \in I_j$. As $|B_{v,i,j}| \leq \mu^2$ and each vertex in $B_{v,i,j}$ has at most $\max\{\lambda, \mu\}$ many common neighbors with z, we have

$$\big|N(z) \cap C_{v,i,j}\big| \leq \mu^2 \cdot \max\{\lambda, \mu\} = O\left(\mu^2 \cdot \frac{k^{4/3}}{n^{1/3}}\right) = o(k)$$

Since $\big|N(z) \cap N^+(f_2(j))\big| = \lambda + 1$, we have the size of each set in the family above is at least

$$k - o(k) - \lambda - 1 = k\big(1 - o(1)\big) = \Omega(k)$$

Since $W_{(u,v),j}$ is a subset of $N(f_2(j))$, every vertex $w \notin N^+(f_2(j))$ can appear in at most μ many $N(z)$'s, $z \in W_{(u,v),j}$, which in turn implies that, if we take the union of all the $|W_{(u,v),j}|$ sets in the family, every vertex can be counted for at most μ times. As a result, we get the following lower bound for $|C_{u,j}|$:

$$|C_{u,j}| \geq |W_{(u,v),j}| \cdot \Omega(k)/\mu = \Omega(\mu \alpha k)$$

The lemma then follows. □

Lemma 4.3.13 (A Lot of Chances for Interaction)**.** $|\cup_{j \in J} C_{u,j}| = \Omega(\mu \alpha \beta k)$.

Proof. For each $j \in J$, we show that $|C_{u,j} - \cup_{j' \neq j} C_{u,j'}| = \Omega(\mu \alpha k)$. The lemma follows by $|J| = \Omega(\beta)$.

To this end, we examine $|C_{u,j} \cap C_{u,j'}|$, for some $j' \in J$ with $j' \neq j$. Note that by the definition of J, $f_2(j)$ and $f_2(j')$ are not connected. Let $W = \cup_{i \in I_j} B_{u,i,j}$ and $W' = \cup_{i \in I_{j'}} B_{u,i,j'}$. Then it is clear that W and W' are subsets of $N(f_2(j))$ and $N(f_2(j'))$, respectively.

We will prove the following inequality:

$$|N(W) \cap N(W')| \leq \mu k + \alpha^2\mu^5 + \alpha\mu^3\lambda \tag{4.21}$$

which gives an upper bound for $|C_{u,j} \cap C_{u,j'}|$ as $C_{u,j} \subseteq N(W)$ and $C_{u,j'} \subseteq N(W')$. To prove (4.21), let

$$W^* = N(f_2(j)) \cap N(f_2(j'))$$

For each vertex $z \in N(W) \cap N(W')$, at least one of the following three cases occurs:

Case 1: $\exists\, w \in W^*$ such that $(w, z) \in E$;

Case 2: $\exists\, w \in N(W - W^*)$ and $w' \in N(W' - W^*)$ such that $(w, w'), (w, z), (w', z) \in E$;

Case 3: $\exists\, w \in N(W - W^*)$ and $w' \in N(W' - W^*)$ such that $(w, z), (w', z) \in E$ and $(w, w') \notin E$.

We can bound the number of z's in each of the three cases respectively by

$$\mu k, \quad (\mu^2\alpha)\mu\lambda, \quad \text{and} \quad (\mu^2\alpha)^2\mu$$

from which (4.21) follows. Using this upper bound for $|C_{u,j} \cap C_{u,j'}|$, we have

$$\left|C_{u,j} - \cup_{j' \neq j} C_{u,j'}\right| \geq |C_{u,j}| - \beta\left(\mu k + \alpha^2\mu^5 + \alpha\mu^3\lambda\right) = \Omega(\mu\alpha k)$$

The lemma then follows from $|J| = \Omega(\beta)$. □

Finally, we prove Theorem 4.3.4:

Proof of Theorem 4.3.4. First of all, with probability $\geq 1 - \exp(-\Omega(\alpha)) - \exp(-\Omega(\beta)) = 1 - \exp(-\Omega(\beta))$, both events described in Lemma 4.3.6 and Lemma 4.3.11 occur. We use $I \subseteq [\alpha]$, $R \subseteq I \times [\beta]$ and $J \subseteq [\beta]$ to denote the three sets that satisfy the conditions of Lemma 4.3.6 and Lemma 4.3.11. Then, whenever

$$f_3(\ell) \in \cup_{j \in J} C_{u,j}$$

for some $\ell \in [\gamma]$, the output of **Cert** would be a triple instead of nil. By Lemma 4.3.13, we have

$$|\cup_{j \in J} C_{u,j}| = \Omega(\mu\alpha\beta k) = \Omega\left(\frac{n^{11/3}}{k^{17/3} \cdot \log^3 n}\right) = \Omega\left(\frac{n}{\gamma} \cdot \log^2 n\right)$$

As a result, the probability that none of the γ vertices $f_3(1), \ldots, f_3(\gamma)$ hits $\cup_{j \in J} C_{u,j}$ is

$$\left(1 - \Omega\left(\frac{\log^2 n}{\gamma}\right)\right)^{\gamma} = \exp\left(-\Omega\left(\log^2 n\right)\right)$$

The theorem then follows from the union bound and the fact that $\beta = \Omega(\log^{7/3} n)$. □

Finally, we prove Theorem 4.3.1.

Proof of Theorem 4.3.1. If $k \geq n/\log n$, by Theorem 4.2.1, $|\operatorname{Aut}(G)| \leq \exp(O(\log^2 n))$. In the following, we assume $k < n/\log n$.

If G does not satisfy claw bound, then by Proposition 4.3.2, Theorem 3.2.2 and 4.3.3, $|\operatorname{Aut}(G)| \leq \exp(O(\log^2 n))$.

Now we assume G satisfies claw bound. By Theorem 4.3.4, if $k \leq n^{7/13}/\log n$, then $|\operatorname{Aut}(G)| = \exp(\widetilde{O}(n^{9/37}))$. By Theorem 4.2.2 and 4.2.5, if $k > n^{7/13}/\log n$, then $|\operatorname{Aut}(G)| = \exp(\widetilde{O}(\sqrt{n/k})) = \exp(\widetilde{O}(n^{3/13}))$. □

Chapter 5

Isomorphism of primitive coherent configurations

A *configuration* $\mathfrak{X}$ on vertex set V is a partition $R_0 \cup \cdots \cup R_{r-1}$ of $V \times V$ with the following properties:

(1) the diagonal $\Delta = \{(v, v) : v \in V\}$ is the union of some of the R_i;

(2) $(\forall i)(\exists i^*)(R_i^{-1} = R_{i^*})$ (where $R_i^{-1} = \{(v, u) : (u, v) \in R_i\}$)

We think of $\mathfrak{X}$ as an edge-colored complete digraph with loops on V, with edge color classes given by the R_i. Hence, the color of a pair $(u, v) \in V \times V$ is $c(u, v) = i$ if $(u, v) \in R_i$. The *rank* r of a configuration is the number of edge color classes. We shall also speak of the colors of the vertices, defined as $c(u) := c(u, u)$. We call the digraph $\mathfrak{X}_i = (V, R_i)$ the *color-i constituent digraph.*

Given a graph $G = (V, E)$, we associate with G the configuration $\mathfrak{X}(G) = (V; \Delta, E, \overline{E})$ where $\overline{E}$ denotes the set of edges of the complement of G. (We omit E if $E = \emptyset$ and omit $\overline{E}$ if $\overline{E} = \emptyset$.) So graphs can be viewed as configurations of rank ≤ 3.

A *coherent configuration* (CC)is a configuration which additionally satisfies the following condition:

(3) for every $0 \leq i, j, k \leq r - 1$, there is a number p_{jk}^i such that for every $(u, v) \in R_i$, there are exactly p_{jk}^i vertices $w \in V$ such that $(u, w) \in R_j$ and $(w, v) \in R_k$.

The numbers p_{jk}^i are called the *structure constants* of $\mathfrak{X}$.

The term "coherent configuration" was coined by Donald Higman in 1969 [Higman, 1970], and at the same time, the same object under a different name was defined by Weisfeiler and Leman [Weisfeiler and Leman, 1968]. In the case corresponding to a permutation group, CCs already effectively appeared in Schur's 1933 paper [Schur, 1933]. This group-theoretic perspective on CCs was developed further by Wielandt [Wielandt, 1964].

A CC is *primitive* (PCC) if it has the following additional properties:

(4) $\Delta = R_0$;

(5) the constituent digraphs $\mathfrak{X}_i$ are strongly connected for every $1 \leq i \leq r-1$.

Given a graph H, the *line-graph* $L(H)$ has as vertices the edges of H, with two vertices adjacent in $L(H)$ if the corresponding edges are incident in H. The *triangular graph* $T(m)$ is the line-graph of the complete graph K_m (so $n = \binom{m}{2}$). The *lattice graph* $L_2(m)$ is the line-graph of the complete bipartite graph $K_{m,m}$ (on equal parts) (so $n = m^2$). Both $T(m)$ and $L_2(m)$ have $\exp(\Omega(m))$ automorphisms.

We say a PCC is *exceptional* if it is of the form $\mathfrak{X}(G)$, where G is isomorphic to the complete graph K_n, the triangular graph $T(m)$, or the lattice graph $L_2(m)$, or the complement of such a graph.

Theorem 5.0.1. *Given a non-exceptional PCC $\mathfrak{X}$, there exists a set of $\widetilde{O}(n^{1/3})$ vertices that completely splits $\mathfrak{X}$ under naive vertex refinement.*

By Lemma 2.2.1,

Corollary 5.0.2. *Let $\mathfrak{X}$ be a non-exceptional PCC $\mathfrak{X}$ with n vertices. We have $|\operatorname{Aut}(\mathfrak{X})| \leq \exp(\widetilde{O}(n^{1/3}))$.*

We remark that it is easy to recognize an exceptional PCC from its clique structure and create a canonical form in polynomial time. Hence

Corollary 5.0.3. *A canonical form of primitive coherent configurations (PCCs) with n vertices can be computed in time $\exp(\widetilde{O}(n^{1/3}))$. In particular, isomorphism of PCCs can be tested within the same time bound.*

In the rest of this chapter, we prove Theorem 5.0.1.

5.1 Growth of spheres

Throughout the chapter, $\mathfrak{X}$ will denote a PCC of rank r on vertex set V with structure constants p^i_{jk} for $0 \le i, j, k \le r-1$. We assume throughout that $r > 2$, since the case $r = 2$ is the trivial case of $\mathfrak{X}(K_n)$, listed as one of our exceptional PCCs.

For any color i in a PCC, we write $n_i = n_{i^*} = p^0_{ii^*} = p^0_{i^*i}$, the out-degree of each vertex in $\mathfrak{X}_i$.

Two colors, 0 and 1, will play a special role. Recall that $R_0 = \Delta$ is the diagonal. Without loss of generality, we assume throughout that $n_1 = \max_i n_i$. We write $\rho = \sum_{i \ge 2} n_i = n - n_1 - 1$.

We say that color 1 is *dominant* if $n_1 \ge n/2$, i.e., $\rho < n/2$. We call a pair of distinct vertices dominant (nondominant) when its color is dominant (nondominant, resp.). We say color i is *symmetric* if $i^* = i$. Note that when color 1 is dominant, it is symmetric, since $n_{1^*} = n_1 \ge n/2$.

For a color i and vertex u, we denote by $\mathfrak{X}_i(u)$ the set of vertices v such that $c(u, v) = i$. We write $N(u)$ for the set of neighbors of u in the graph $G(\mathfrak{X})$. For i nondominant, we define $\lambda_i = |\mathfrak{X}_i(u) \cap N(v)|$, where $c(u, v) = i$. So, the parameters λ_i are loosely analogous to the parameter λ of a SRG.

For a nondominant color i and vertex u, the *δ-sphere* $\mathfrak{X}_i^{(\delta)}(u)$ in $\mathfrak{X}_i$ centered at u is the set of vertices v with $\mathrm{dist}_i(u, v) = \delta$.

We prove

Lemma 5.1.1 (Growth of spheres)**.** *Let $\mathfrak{X}$ be a PCC, let $i, j \ge 1$ be nondiagonal colors, let $\delta = \mathrm{dist}_i(j)$, and $u \in V$. Then for any integer $1 \le \alpha \le \delta - 2$, we have*

$$|\mathfrak{X}_i^{(\alpha+1)}(u)||\mathfrak{X}_i^{(\delta-\alpha)}(u)| \ge n_i n_j.$$

We note that Lemma 5.1.1 is straightforward when $\mathfrak{X}_i$ is distance-regular. Indeed, a significant portion of the difficulty of the lemma was in finding the correct generalization.

We will use Lemma 5.1.1 to prove Lemma 5.2.1 below, which shows that a modest number of individualizations suffice to completely split $\mathfrak{X}$ when ρ is sufficiently large. We thereby reduce to the case that $\rho = o(n^{2/3})$.

We start from a few basic observations.

Proposition 5.1.2. *Let $G = (A, B, E)$ be a bipartite graph, and let $A_1 \cup \cdots \cup A_m$ be a partition of A such that the subgraph induced on (A_i, B) is biregular of positive valency for each $1 \le i \le m$. Then for any $A' \subseteq A$, we have*

$$|N(A')|/|A'| \ge |B|/|A|$$

where $N(A')$ is the set of neighbors of vertices in A', i.e., $N(A') = \{y \in B : \exists x \in A', \{x, y\} \in E\}$.

Proof. Let $A' \subseteq A$. By the pigeonhole principle, there is some i such that $|A' \cap A_i|/|A_i| \ge |A'|/|A|$. Let α be the degree of a vertex in A_i and let β be the number of neighbors in A_i of a vertex in B. We have $\alpha|A_i| = \beta|B|$, and $\beta|N(A' \cap A_i)| \ge \alpha|A' \cap A_i|$. Hence,

$$|N(A')| \ge |N(A' \cap A_i)| \ge \frac{|A' \cap A_i|\alpha}{\beta} = \frac{|A' \cap A_i||B|}{|A_i|} \ge \frac{|B||A'|}{|A|}. \qquad \square$$

Suppose $A, B \subseteq V$ are disjoint set of vertices. We denote by (A, B, i) the bipartite graph between A and B such that there is an edge from $x \in A$ to $y \in B$ if $c(x, y) = i$. For $I \subseteq [r-1]$ a set of nondiagonal colors, we denote by (A, B, I) the bipartite graph between A and B such that there is an edge from $x \in A$ to $y \in B$ if $c(x, y) \in I$.

Fact 5.1.3. *For any vertex u, colors $0 \le j, k \le r-1$ with $j \ne k$, and set $I \subseteq [r-1]$ of nondiagonal colors, the bipartite graph $(\mathfrak{X}_j(u), \mathfrak{X}_k(u), I)$ is biregular.*

Proof. The degree of every vertex in $\mathfrak{X}_j(u)$ is $\sum_{i \in I} p^j_{ik^*}$. And the degree of every vertex in $\mathfrak{X}_k(u)$ is $\sum_{i \in I} p^k_{ji}$. $\square$

Recall our notation $\mathfrak{X}_i^{(\delta)}(u)$ for the δ-sphere centered at u in the color-i constituent digraph, i.e., the set of vertices v such that $\mathrm{dist}_i(u, v) = \delta$.

For the remainder of Section 5.1, we fix a PCC $\mathfrak{X}$, a color $1 \le i \le r-1$, and a vertex u. For a color $1 \le j \le r-1$ and an integer $1 \le \alpha \le \mathrm{dist}_i(j)$, we denote by $S_\alpha^{(j)}$ the set of vertices $v \in \mathfrak{X}_i^{(\alpha)}(u)$ such that there is a vertex $w \in \mathfrak{X}_j(u)$ and a shortest path in $\mathfrak{X}_i$ from u to w passing through v, i.e.,

$$S_\alpha^{(j)} = \{v \in \mathfrak{X}_i^{(\alpha)}(u) : \exists w \in \mathfrak{X}_j(u) \text{ s.t. } \mathrm{dist}_i(u, v) + \mathrm{dist}_i(v, w) = \mathrm{dist}_i(u, w)\}.$$

Note that these sets $S_\alpha^{(j)}$ are nonempty by the primitivity of $\mathfrak{X}$, and in particular, if $\alpha = \mathrm{dist}_i(j)$, then $S_\alpha^{(j)} = \mathfrak{X}_j(u)$. For $v \in V$ and an integer $\mathrm{dist}_i(u, v) < \alpha \le \mathrm{dist}_i(j)$, we denote

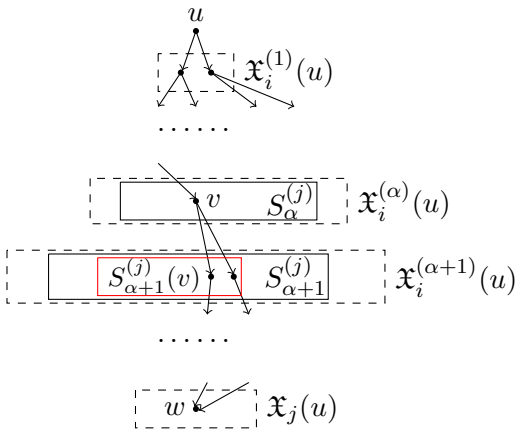

Figure 5.1: Growth of spheres for primitive coherent configurations

by $S_\alpha^{(j)}(v) \subseteq S_\alpha^{(j)}$ the set of vertices $x \in S_\alpha^{(j)}$ such that there is a shortest path in $\mathfrak{X}_i$ from u to x passing through v, i.e.

$$S_\alpha^{(j)}(v) = S_\alpha^{(j)} \cap \mathfrak{X}_i^{(\alpha - \mathrm{dist}_i(u,v))}(v) = \{x \in S_\alpha^{(j)} : \mathrm{dist}_i(u,v) + \mathrm{dist}_i(v,x) = \mathrm{dist}_i(u,x)\}.$$

See Figure 1 for a graphical explanation of the notation.

Corollary 5.1.4. *Let $1 \le j \le r-1$ be a color such that $\delta = \mathrm{dist}_i(j) \ge 3$. Let $1 \le \alpha \le \delta - 2$ be an integer, and let $v \in S_\alpha^{(j)}$. Then*

$$\frac{|S_\delta^{(j)}(v)|}{|S_{\alpha+1}^{(j)}(v)|} \ge \frac{n_j}{|S_{\alpha+1}^{(j)}|}.$$

Proof. Consider the bipartite graph $(S_{\alpha+1}^{(j)}, \mathfrak{X}_j(u), I)$ with

$$I = \{k : 1 \le k \le r-1 \text{ and } \mathrm{dist}_i(k) = \mathrm{dist}_i(j) - \alpha - 1\}.$$

There is an edge from $x \in S_{\alpha+1}^{(j)}$ to $y \in \mathfrak{X}_j(u)$ if there is a shortest path from u to y passing through x.

By the coherence of $\mathfrak{X}$, if $\mathfrak{X}_\ell(u) \cap S_{\alpha+1}^{(j)}$ is nonempty for some color ℓ, then $\mathfrak{X}_\ell(u) \subseteq S_{\alpha+1}^{(j)}$. Hence, $S_{\alpha+1}^{(j)}$ is partitioned into sets of the form $\mathfrak{X}_\ell(u)$ with $\mathrm{dist}_i(\ell) = \alpha + 1$. For such colors ℓ, by Fact 5.1.3, $(\mathfrak{X}_\ell(u), \mathfrak{X}_j(u), I)$ is biregular, and by the definition of $S_{\alpha+1}^{(j)}$, then $(\mathfrak{X}_\ell(u), \mathfrak{X}_j(u), I)$ is not an empty graph.

Therefore, the result follows by applying Proposition 5.1.2 with $A = S^{(j)}_{\alpha+1}$, $B = \mathfrak{X}_j(u)$, $A' = S^{(j)}_{\alpha+1}(v) \subseteq S^{(j)}_{\alpha+1}$, and (hence) $N(A') = S^{(j)}_{\delta}(v)$. □

Fact 5.1.5. *Let $1 \le j \le r-1$ be a color such that $\delta = \mathrm{dist}_i(j) \ge 3$, and w be a vertex in $\mathfrak{X}_j(u)$. Let $1 \le \alpha \le \delta - 2$, and let v be a vertex in $S^{(j)}_{\alpha}$. If $\mathrm{dist}_i(v,w) = \delta - \alpha$, then*

$$\{x : x \in \mathfrak{X}_i(v) \text{ and } \mathrm{dist}_i(x,w) = \delta - \alpha - 1\} \subseteq S^{(j)}_{\alpha+1}(v).$$

Proof. For any $x \in \mathfrak{X}_i(v)$, we have $\mathrm{dist}_i(u,x) \le \alpha + 1$. If $\mathrm{dist}_i(x,w) = \delta - \alpha - 1$, then $x \in \mathfrak{X}^{(\alpha+1)}_i(u)$, because otherwise $\mathrm{dist}(u,w) < \delta$. Then x is in $S^{(j)}_{\alpha+1}(v)$, since there is a shortest from u to w passing through x. □

Proposition 5.1.6. *Let $1 \le j \le r-1$ be a color such that $\delta = \mathrm{dist}_i(j) \ge 3$. Let $1 \le \alpha \le \delta - 2$, and let $v \in S^{(j)}_{\alpha}$. Then*

$$|\mathfrak{X}^{\delta-\alpha}_i(u)| \ge \frac{n_i |S^{(j)}_{\delta}(v)|}{|S^{(j)}_{\alpha+1}(v)|}.$$

Proof. Let k be a color satisfying $\mathrm{dist}_i(k) = \delta - \alpha$ and $\mathfrak{X}_k(v) \cap S^{(j)}_{\delta}(v) \neq \emptyset$. Let w be a vertex in $\mathfrak{X}_k(v) \cap S^{(j)}_{\delta}(v)$. Consider the bipartite graph $B = (\mathfrak{X}_i(v), \mathfrak{X}_k(v), I)$, where $I = \{\ell : \mathrm{dist}_i(\ell) = \delta - \alpha - 1\}$.

By Fact 5.1.3, B is biregular, and by Fact 5.1.5 the degree of w in B is at most $|S^{(j)}_{\alpha+1}(v)|$. Denote by d_k the degree of a vertex $x \in \mathfrak{X}_i(v)$ in B, so $n_k |S^{(j)}_{\alpha+1}(v)| \ge n_i d_k$. Hence, summing over all colors k such that $\mathfrak{X}_k(v) \cap S^{(j)}_{\delta}(v) \neq \emptyset$, we have

$$|\mathfrak{X}^{(\delta-\alpha)}_i(v)| \ge \sum_k n_k \ge \sum_k \frac{n_i d_k}{|S^{(j)}_{\alpha+1}(v)|} \ge \frac{n_i |S^{(j)}_{\delta}(v)|}{|S^{(j)}_{\alpha+1}(v)|}.$$

Finally, by the coherence of $\mathfrak{X}$, we have $|\mathfrak{X}^{(\delta-\alpha)}_i(u)| = |\mathfrak{X}^{(\delta-\alpha)}_i(v)|$. □

We now complete the proof of Lemma 5.1.1.

Proof of Lemma 5.1.1. Combining Corollary 5.1.4 and Proposition 5.1.6, for any $1 \le \alpha \le \delta - 2$ we have

$$|\mathfrak{X}^{(\delta-\alpha)}_i(u)| \ge \frac{n_i n_k}{|S^{(k)}_{\alpha+1}|}$$

and so since $S^{(k)}_{\alpha+1} \subseteq \mathfrak{X}^{(\alpha+1)}_i(u)$ by definition, we have the desired inequality. □

5.2 Distinguishing number

In this section, we will prove following lemma, which will allow us to assume that our PCCs $\mathfrak{X}$ satisfy $\rho = o(n^{2/3})$.

Lemma 5.2.1. *Let $\mathfrak{X}$ be a PCC. If $\rho \geq n^{2/3}(\log n)^{-1/3}$, then there is a set of vertices with size $O(n^{1/3}(\log n)^{4/3})$ which completely splits $\mathfrak{X}$ under naive vertex refinement.*

Following Babai [Babai, 1981b], we analyze the distinguishing number.

Definition 5.2.2. *Let $u, v \in V$. We say $w \in V$ **distinguishes** u and v if $c(w,u) \neq c(w,v)$. We write $D(u,v)$ for the set of vertices w distinguishing u and v, and $D(i) = |D(u,v)|$ where $c(u,v) = i$. We call $D(i)$ the **distinguishing number** of i.*

Hence, $D(i) = \sum_{j \neq k} p^i_{jk^*}$. If $w \in D(u,v)$, then after individualizing w and refining, u and v get different colors.

Lemma 5.2.3 ([Babai, 1981b, Lemma 5.4])**.** *Let $\mathfrak{X}$ be a PCC and let $\zeta = \min\{D(i) : 1 \leq i \leq r-1\}$. Then there is a set of size $O((n \log n)/\zeta)$ which completely splits $\mathfrak{X}$ under naive vertex refinement.*

Thus, to prove Lemma 5.2.1, we show that if $\rho \geq n^{2/3}(\log n)^{-1/3}$ then for every color $i \neq 0$, we have $D(i) = \Omega(n^{2/3}(\log n)^{-1/3})$.

We give the following lower bound on ζ when ρ is sufficiently large.

Lemma 5.2.4. *Let $\mathfrak{X}$ be a PCC. If $\rho \geq n^{2/3}(\log n)^{-1/3}$, then $D(i) = \Omega(n^{2/3}(\log n)^{-1/3})$ for all $1 \leq i \leq r-1$.*

Lemma 5.2.1 follows immediately from Lemmas 5.2.3 and 5.2.4. □

We will prove Lemma 5.2.4 by separately addressing the cases $\rho \geq n/3$ and $\rho < n/3$. The case $\rho < n/3$ will rely on our estimate for the size of spheres in constituent digraphs, Lemma 5.1.1. For the case $\rho \geq n/3$, we will rely on following lemma.

Lemma 5.2.5. *Let $\mathfrak{X}$ be a PCC. For any nondiagonal color i, the number of colors j such that $n_j > n_i/2$ is at most $O((\log n + n/\rho)D(i)/n_i)$.*

We first recall the following observations of Babai [Babai, 1981b, Proposition 6.3].

Proposition 5.2.6 (Babai)**.** *Let $\mathfrak{X}$ be a PCC. Then*

$$\frac{1}{n-1}\sum_{j=1}^{r-1} D(j)n_j \geq \rho + 2.$$

The following corollary is then immediate. □

Corollary 5.2.7. *Let $\mathfrak{X}$ be a PCC. There exists a nondiagonal color i with $D(i) > \rho$.*

The following facts about the parameters of a coherent configuration are standard.

Proposition 5.2.8 ([Zieschang, 2010, Lemma 1.1.1, 1.1.2, 1.1.3])**.** *Let $\mathfrak{X}$ be a CC. Then for all colors i, j, k, the following relations hold:*

1. $n_i = n_{i^*}$
2. $p^i_{jk} = p^{i^*}_{k^*j^*}$
3. $n_i p^i_{jk} = n_j p^j_{ik^*}$
4. $\sum_{j=0}^{r-1} p^i_{jk} = \sum_{j=0}^{r-1} p^i_{kj} = n_k$

5.2.1 Bound on the number of large colors

We now prove Lemma 5.2.5, using the following preliminary results.

Lemma 5.2.9. *Let $\mathfrak{X}$ be a PCC, let I be a nonempty set of nondiagonal colors, let $n_I = \sum_{i\in I} n_i$, and let J be the set of colors j such that $n_j \leq n_I/2$. Then*

$$\sum_{j\in J} n_j \leq 2\max\{D(i) : i \in I\}.$$

Proof. For any color i, by Proposition 5.2.8, we have

$$\begin{aligned} D(i) &= \sum_{j=0}^{r-1}\sum_{k\neq j} p^i_{jk^*} = \sum_{j=0}^{r-1}\sum_{k\neq j} \frac{n_j p^j_{ik}}{n_i} \\ &= \frac{1}{n_i}\sum_{j=0}^{r-1} n_j \sum_{k\neq j} p^j_{ik} = \frac{1}{n_i}\sum_{j=0}^{r-1} n_j(n_i - p^j_{ij}). \end{aligned}$$

Therefore,

$$\begin{aligned} n_I \max\{D(i) : i \in I\} &\geq \sum_{i\in I} n_i D(i) \\ &\geq \sum_{i\in I}\sum_{j\in J} n_j(n_i - p_{ij}^j) \\ &\geq \sum_{j\in J} n_j \sum_{i\in I}(n_i - p_{ij}^j) \\ &\geq \sum_{j\in J} n_j\,(n_I - n_j) \\ &\geq \frac{n_I}{2}\left(\sum_{j\in J} n_j\right). \end{aligned}$$

□

Lemma 5.2.10. *Let $\mathfrak{X}$ be a PCC, and suppose $p_{jk}^i > 0$ for some i, j, k. Then*

$$D(j) - D(k) \leq D(i) \leq D(j) + D(k).$$

Proof. Fix vertices $u, v, w \in V$ with $c(u, w) = i$, $c(u, v) = j$, and $c(v, w) = k$. (These vertices exist since $p_{jk}^i > 0$.) For any vertex x such that $c(x, u) \neq c(x, w)$, we have $c(x, u) \neq c(x, v)$ or $c(x, v) \neq c(x, w)$. Therefore, $D(j) + D(k) \geq D(i)$.

For the other inequality, if $p_{jk}^i > 0$ then $p_{ik^*}^j > 0$ by Proposition 5.2.8, and $D(k^*) = D(k)$ by the definition of distinguishing number. So we have $D(i)+D(k) = D(i)+D(k^*) \geq D(j)$, using the previous paragraph for the latter inequality. □

Lemma 5.2.11. *Let $\mathfrak{X}$ be a PCC. Then for any nondiagonal color i and number $0 \leq \eta \leq \rho - D(i)$, there is a color j such that $\eta < D(j) \leq \eta + D(i)$.*

Proof. By Corollary 5.2.7, there is a color k with $D(k) > \rho$. Now consider a shortest path $u_0, \ldots, u_\ell$ in $\mathfrak{X}_i$ with $c(u_0, u_\ell) = k$. (By the primitivity of $\mathfrak{X}$, the digraph $\mathfrak{X}_i$ is strongly connected, and such a path exists.) Let $\delta_j = D(c(u_0, u_j))$ for $1 \leq j \leq k$. By Lemma 5.2.10, we have $|\delta_j - \delta_{j+1}| \leq D(i)$. Hence, one of the numbers δ_j falls in the interval $(\eta, \eta + D(i)]$ for any $0 \leq \eta \leq \rho - D(i)$. □

We denote by I_α the set of colors i with $D(i) \leq \alpha$.

Lemma 5.2.12. *Let $\mathfrak{X}$ be a PCC with $\rho > 0$. Let i be a nondiagonal color and let $0 \le \eta \le \rho - 2D(i)$. Then*

$$n_i \le \sum_{j \in I_{\eta+3D(i)} \setminus I_\eta} n_j.$$

Proof. By Lemma 5.2.11, the set $I_{\eta+2D(i)} \setminus I_{\eta+D(i)}$ is nonempty. Let $k \in I_{\eta+2D(i)} \setminus I_{\eta+D(i)}$. We have $\sum_{j=0}^{r-1} p_{ij}^k = n_i$ by Proposition 5.2.8. On the other hand, if $p_{ij}^k > 0$ for some j, then $D(j) - D(i) \le D(k) \le D(j) + D(i)$ by Lemma 5.2.10, and so $j \in I_{\eta+3D(i)} \setminus I_\eta$. Hence,

$$n_i = \sum_{j=0}^{r-1} p_{ij}^k = \sum_{j \in I_{\eta+3D(i)} \setminus I_\eta} p_{ij}^k \le \sum_{j \in I_{\eta+3D(i)} \setminus I_\eta} n_j.$$

□

Lemma 5.2.13. *Let $\mathfrak{X}$ be a PCC with $\rho > 0$, let i be a nondiagonal color, and let $0 \le \eta \le \rho$. Then*

$$\left\lfloor \frac{\eta}{3D(i)} \right\rfloor n_i \le \sum_{j \in I_\eta} n_j.$$

Proof. If $\eta < 3D(i)$, the left-hand side is 0, so assume $\eta \ge 3D(i)$. For any integer $1 \le \alpha \le \lfloor \eta/(3D(i)) \rfloor$, let $S_\alpha = I_{3D(i)\alpha} \setminus I_{3D(i)(\alpha-1)}$. Then

$$\bigcup_{\alpha=1}^{\lfloor \eta/(3D(i)) \rfloor} S_\alpha \subseteq I_\eta$$

By the disjointness of the sets S_α and Lemma 5.2.12, we have

$$\sum_{j \in I_\eta} n_j \ge \sum_{\alpha=1}^{\lfloor \eta/(3D(i)) \rfloor} \sum_{j \in S_\alpha} n_j \ge \left\lfloor \frac{\eta}{3D(i)} \right\rfloor n_i.$$

□

Finally, we are able to prove Lemma 5.2.5.

Proof of Lemma 5.2.5. Fix an integer $0 \le \alpha \le \lfloor \log_2(\rho/(3D(i))) \rfloor$. For any number β, let J_β denote the set of colors j such that $n_j \le \beta$. We start by estimating $|J_{2^\alpha n_i} \setminus J_{2^{\alpha-1} n_i}|$, i.e., the number of colors j with $2^{\alpha-1} n_i < n_j \le 2^\alpha n_i$. By Lemma 5.2.13, we have

$$\sum_{j \in I_{2^\alpha(3D(i))}} n_j \ge 2^\alpha n_i.$$

Therefore, applying Lemma 5.2.9 with $I = I_{2^\alpha(3D(i))}$ and $J = J_{2^\alpha n_i}$, we have

$$\sum_{j \in J_{2^\alpha n_i}} n_j \le 2 \max\{D(i) : i \in I_{2^\alpha \cdot 3D(i)}\} \le 2^{\alpha+1}(3D(i)),$$

with the second inequality coming from the definition of $I_{2^\alpha(3D(i))}$.

It follows that the number of colors j such that $j \in J_{2^\alpha n_i} \setminus J_{2^{\alpha-1} n_i}$ is at most

$$2^{\alpha+1}(3D(i))/(2^{\alpha-1}n_i) = 12D(i)/n_i.$$

Overall, the number of colors j satisfying

$$(1/2)n_i < n_j \le 2^{\lfloor \log_2(\rho/3D(i)) \rfloor} n_i$$

is at most $12(\log_2 n + 1)D(i)/n_i$.

Furthermore, the number of colors j satisfying

$$n_j > 2^{\lfloor \log_2(\rho/3D(i)) \rfloor} n_i \ge \frac{\rho n_i}{6D(i)}$$

is at most $(6D(i)/(\rho n_i))n$, since $\sum_{j=0}^{r-1} n_j = n$. Hence, there are at most $O((\log n + n/\rho)D(i)/n_i)$ colors j such that $n_j > n_i/2$. □

5.2.2 Estimates of the distinguishing number

We now prove Lemma 5.2.4, our lower bound for $D(i)$.

First, we recall the following two observations made by Babai [Babai, 1981b, Proposition 6.4 and Theorem 6.11].

Proposition 5.2.14 (Babai)**.** *Let $\mathfrak{X}$ be a PCC. For colors $0 \le i, j \le r-1$, we have*

$$D(j) \le \mathrm{dist}_i(j) D(i).$$

Proposition 5.2.15 (Babai)**.** *Let $\mathfrak{X}$ be a PCC. For any color $1 \le i \le r-1$, we have $n_i D(i) \ge n - 1$.*

We prove the following two estimates of the distinguish number.

Lemma 5.2.16. *Let $\mathfrak{X}$ be a PCC. Fix nondiagonal colors $i, j \ge 1$ and a vertex $u \in V$. Let $\delta = \mathrm{dist}_i(j)$, and $\gamma = \sum_{\alpha=2}^{\delta-1} |\mathfrak{X}_i^{(\alpha)}(u)|$. If $\delta \ge 3$, then*

$$D(i) = \Omega\left(\left(\frac{D(j)\sqrt{n n_j}}{\gamma}\right)^{2/3}\right).$$

Proof. By Lemma 5.1.1, for any $1 \le \alpha \le \delta - 2$ we have

$$|\mathfrak{X}_i^{(\alpha+1)}(u)||\mathfrak{X}_i^{(\delta-\alpha)}(u)| \ge n_i n_j$$

and in particular,

$$\max\{|\mathfrak{X}_i^{(\alpha+1)}(u)|, |\mathfrak{X}_i^{(\delta-\alpha)}(u)|\} \ge \sqrt{n_i n_j}.$$

Hence,

$$\gamma = \sum_{\alpha=2}^{\delta-1} |\mathfrak{X}_i^{(\alpha)}(u)| = \Omega(\delta\sqrt{n_i n_j}) = \Omega\left(\frac{\delta\sqrt{n n_j}}{\sqrt{D(i)}}\right), \tag{5.1}$$

where the last inequality comes from Proposition 5.2.15. Now by Proposition 5.2.14 and Eq. (5.1), we have

$$D(i) \ge \frac{D(j)}{\delta} = \Omega\left(\frac{D(j)\sqrt{n n_j}}{\gamma\sqrt{D(i)}}\right),$$

from which the desired inequality immediately follows. □

Lemma 5.2.17. *Let $\mathfrak{X}$ be a PCC with $\rho = \Omega(n)$. Then every nondiagonal color i with $n_i \le \rho$ satisfies*

$$D(i) = \Omega\left(\sqrt{\frac{\rho n_i}{\log n}}\right).$$

Proof. Fix a nondiagonal color i with $n_i \le \rho$, and suppose $D(i) < \rho/6$ (otherwise the lemma holds trivially). Let J_β denote the set of colors j such that $n_j \le \beta$. Applying Lemma 5.2.9 with the set $I = \{i\}$, we have

$$\sum_{j \in J_{n_i/2}} n_j \le 2D(i). \tag{5.2}$$

On the other hand, by Lemma 5.2.12, for every integer η with $0 \le \eta \le \rho/2 - 3D(i)$,

$$n_i \le \sum_{j \in I_{\eta+3D(i)} \setminus I_\eta} n_j.$$

Thus, for every such η, at least one of following two conditions hold:

(i) there exists a color $j \in I_{\eta+3D(i)} \setminus I_\eta$ satisfying $n_j > n_i/2$;

(ii) $\displaystyle\sum_{\substack{j \in I_{\eta+3D(i)} \setminus I_\eta: \\ n_j \le n_i/2}} n_j \ge n_i.$

There are at least $\lfloor \rho/(6D(i)) \rfloor$ disjoint sets of the form $I_{\eta+3D(i)} \setminus I_\eta$ with $0 \le \eta \le \rho/2 - 3D(i)$. By Lemma 5.2.5, at most $O((\log n + n/\rho)D(i)/n_i) = O((\log n)D(i)/n_i)$ of these satisfy (i). By Eq. (5.2), at most $2D(i)/n_i$ satisfy (ii). Hence, $\lfloor \rho/(6D(i)) \rfloor = O((\log n)D(i)/n_i)$, giving the desired inequality. □

We recall that when color 1 is dominant, it is symmetric. In this case, we recall our notation $\mu = |N(x) \cap N(y)|$, where $x, y \in V$ are any pair of vertices with $c(x, y) = 1$ and $N(x)$ is the nondominant neighborhood of x. Hence, $\mu = \sum_{i,j>1} p^1_{ij}$.

Lemma 5.2.18. *Let $\mathfrak{X}$ be a PCC with $n_1 \ge n/2$. Then $\mu \le \rho^2/n_1$.*

Proof. Fix a vertex u. There are at most ρ^2 paths of length two from u along edges of nondominant color, and exactly n_1 vertices v such that $c(u, v) = 1$. For any such vertex y, there are exactly μ paths of length two from u to v along edges of nondominant color. Hence, $\mu \le \rho^2/n_1$. □

Proof of Lemma 5.2.4. First, suppose $n^{2/3}(\log n)^{-1/3} \le \rho < n/3$. We have $n_1 = n - \rho - 1 > 2n/3 - 1$. Consider two vertices $u, v \in V$ with $c(u, v) = 1$. Note that for any vertex $w \in N(v) \setminus N(u)$, we have $c(w, u) = 1$ and $c(w, v) > 1$. Hence, by Lemma 5.2.18 and the definition of $D(1)$,

$$D(1) \ge \rho - \mu \ge \rho - \frac{\rho^2}{n_1} \ge \left(\frac{1}{2} - o(1)\right)\rho = \Omega(n^{2/3}(\log n)^{-1/3}).$$

Fix a nondominant color i. If $\mathrm{dist}_i(1) = 2$, then by Proposition 5.2.14,

$$D(i) \ge \frac{D(1)}{2} \ge \Omega(n^{2/3}(\log n)^{1/3}).$$

Otherwise, if $\mathrm{dist}_i(1) \ge 3$, by applying Lemma 5.2.16 with $j = 1$, we have

$$D(i) = \Omega\left(\left(\frac{D(1)\sqrt{nn_1}}{n - n_1}\right)^{2/3}\right) = \Omega\left(\left(\frac{\rho n}{\rho - 1}\right)^{2/3}\right) = \Omega(n^{2/3}).$$

Now suppose $\rho \ge n/3$. By Lemma 5.2.17 and Proposition 5.2.15, for every color i with $n_i \le \rho$, we have

$$(D(i))^{3/2} = \Omega\left(\sqrt{\frac{\rho n_i D(i)}{\log n}}\right) = \Omega\left(\sqrt{\frac{\rho n}{\log n}}\right),$$

and hence $D(i) = \Omega(n^{2/3}(\log n)^{-1/3})$. If $n_1 \leq \rho$, then $n_i \leq \rho$ for all i, and we are done. Otherwise, if $n_1 > \rho$, we have only to verify that $D(1) = \Omega(n^{2/3}(\log n)^{-1/3})$. Consider two vertices u, w with $\mathrm{dist}_1(u, w) = 2$. (Since we assume the rank is at least 3, we can always find such u, w by the primitivity of $\mathfrak{X}$.) Let $i = c(u, w)$. Then $i > 1$ and so $n_i \leq \rho$. Since $D(i) = \Omega(n^{2/3}(\log n)^{-1/3})$ for every color $1 < i \leq r - 1$, and $\mathrm{dist}_1(i) = 2$, we have $D(1) = \Omega(n^{2/3}(\log n)^{-1/3})$ by Proposition 5.2.14. □

We have now reduced to the case that $\rho = o(n^{2/3})$. Our analysis of this case is inspired by Spielman's analysis of SRGs [Spielman, 1996].

Lemma 5.2.19. *There exists a constant $\varepsilon > 0$ such that the following holds. Let $\mathfrak{X}$ be a PCC with $\rho = o(n^{2/3})$. There is a set of $O(n^{1/4}(\log n)^{1/2})$ vertices which completely splits $\mathfrak{X}$ under naive vertex refinement.*

For a proof of Lemma 5.2.19, see [Sun and Wilmes, 2015] and [Wilmes, 2016]. Combining Lemma 5.2.1 and 5.2.19, we obtain Theorem 5.0.1.

Chapter 6

Property testing of graph isomorphism

In this chapter we study a property testing version of the graph isomorphism problem. We want to distinguish pairs of graphs that are isomorphic from pairs of graphs that are significantly different.

We define the distance of two graphs, $G = (V_G, E_G)$ and $H = (V_H, E_H)$, as the minimum of the normalized Hamming distances of their respective isomorphic copies, i.e., the minimum number of edges that have to be modified in G (added or deleted) to turn it into a graph isomorphic to H, divided by $\binom{n}{2}$. The distance is zero iff they are isomorphic; so this is not a metric on graphs but a metric on isomorphism classes of graphs. We say that two graphs are ε-*far* if their distance is at least ε.

In the property testing version of the graph isomorphism problem, we want an algorithm to accept with probability at least 9/10 if the input graphs are isomorphic and reject with probability at least 9/10 if the input graphs are ε-far, where $\varepsilon > 0$ is the distance parameter that is passed to the algorithm. One or two input graphs are not explicitly given to the algorithm, but the edge query oracle is provided. (The algorithm can query, for example, if the first vertex and the fifth vertex are adjacent in graph G.) The goal of a property testing algorithm is to distinguish the two cases using as few edge queries as possible.

Fischer and Matsliah [Fischer and Matsliah, 2008], who were the first to study the prob-

lem, consider different versions of the problem based on whether both graphs are unknown and whether the algorithm has to satisfy a stronger requirement of never rejecting isomorphic pairs of graphs. We now discuss different flavors of the problem in more detail.

One graph known vs. both graphs unknown: The main goal of property testing research is to determine the fraction of the input the algorithm has to query in order to solve a problem. In this work, we use the dense graph model, in which the algorithm can check in a single query whether an edge connects two arbitrary vertices of an unknown input graph. There are two natural versions of the graph isomorphism problem. In one, the algorithm completely knows a graph G (or alternately, queries about G do not count towards its complexity) and only has to query pairs of vertices of H. In the other, the algorithm initially knows nothing about either of the graphs and has to query edges in both of them. Clearly, the complexity of isomorphism testing in the former model is not higher than in the latter.

One-sided vs. two-sided error: The standard definition of property testing (as above) allows algorithms to err with small constant probability both when the input graphs are isomorphic and when they are ε-far. It may sometimes be desirable to ensure that the algorithm never rejects pairs of graphs that are isomorphic. This kind of setting is referred to as testing with *one-sided error*, compared to the former, which is referred to as testing with *two-sided error*. In the one-sided error setting, if the input graphs are ε-far, the algorithm has to find evidence of their non-isomorphism with high constant probability. In the two-sided error setting, it suffices that the algorithm collects enough information to prove that graphs are unlikely to be isomorphic in order to reject them. The query complexity of two-sided error testing is never higher that the query complexity of one-sided error testing. It is not uncommon that there is a sharp difference between the complexity of the two versions of testing problems. Fischer and Matsliah showed that this is the case for the graph isomorphism problem (see Table 6.1, which we discuss next).

Fischer and Matsliah consider four versions of the problem resulting from combining the above options. Table 6.1 presents their results and our contribution to the understanding

Version of the problem	**Previous results [Fischer and Matsliah, 2008]**	**This work**
One-sided error, one graph known	$\tilde{O}(n)$, $\Omega(n)$	
One-sided error, both graphs unknown	$\tilde{O}\left(n^{3/2}\right)$, $\Omega\left(n^{3/2}\right)$	
Two-sided error, one graph known	$\tilde{O}\left(n^{1/2}\right)$, $\Omega\left(n^{1/2}\right)$	
Two-sided error, both graphs unknown	$\tilde{O}\left(n^{5/4}\right)$, $\Omega\left(n\right)$	$O\left(n\right) \cdot 2^{\widetilde{O}(\sqrt{\log n})}$

Table 6.1: The query complexity of property testing of graph isomorphism

of the problem. Both their and our focus is on the dependence on n, i.e., the number of vertices, once the proximity parameter ε is fixed to a positive constant. They obtained bounds optimal up to polylogarithmic factors in all but one case in which two-sided error is allowed and the algorithm does not know either of the graphs. Arguably, this is the most interesting case. First, small probability of rejecting isomorphic graphs may be acceptable, because we can make it an arbitrarily small constant by repeating the algorithm and taking the majority of answers. Second, if we take the big data perspective, it may be difficult for the algorithm to just "know" one of the graphs and it may make much more sense to assume that the algorithm has to query both of them. Our contribution is essentially closing the gap between $\Omega(n)$ and $\tilde{O}\left(n^{5/4}\right)$ in this case. More precisely, we prove the following theorem

Theorem 6.0.1. *There is an algorithm solving the property testing of graph isomorphism problem for two unknown graphs with two sided error with parameter* $\varepsilon_0 < 1$ *using* $n \cdot \exp\left(O\left(\frac{\log\log n\sqrt{\log n}}{\varepsilon_0}\right)\right)$ *queries with running time* $\exp\left(O\left(\frac{\log\log n(\log n)^{1.5}}{\varepsilon_0}\right)\right)$.

In this chapter, we prove Theorem 6.0.1. We will overview the high level idea of our improvement in Section 6.1, and present the details of the proof in the following sections.

6.1 Overview of the proof

In this section, we first review Fischer and Matsliah's algorithm, and then go over the high level idea of our improvement.

6.1.1 Overview of the paper of Fischer and Matsliah

We briefly sketch the high level idea of the previous upper bound using $\widetilde{O}(n^{5/4})$ samples by Fischer and Matsliah [Fischer and Matsliah, 2008].

Given a graph G. A core set of G is a list of vertices in G. Fixing a core set of k vertices in graph G $(x_1, x_2, \ldots, x_k)$, the label of a vertex x is defined as $L(x) = e(x, x_1) \circ e(x, x_1) \circ \cdots \circ e(x, x_k)$, a binary string of length k whose i-th bit is 1 iff x is adjacent to x_i, otherwise i-th bit is 0.

Theorem 6.1.1 ([Fischer and Matsliah, 2008], restated)**.** *Let C_G and C_H be two core sets of G and H with size $(\log n)^2$. If the following two conditions satisfy*

1. *The label distribution of vertices in G with respect to C_G has total variation distance $\varepsilon/10$ to the label distribution of vertices in H with respect to C_H;*

2. *If the vertices are sampled in the following way: randomly sample two vertices u, v in G, and then randomly sample two vertices u', v' in H satisfying $L(u) = L(u'), L(v) = L(v')$, then the probability of $e_G(u, v) = e_H(u', v')$ is at least $1 - \varepsilon/4$.*

Then G and H is at most ε-far.

The high level idea of [Fischer and Matsliah, 2008] is to find a pair of C_G and C_H satisfying the two conditions of Theorem 6.1.1. The algorithm of [Fischer and Matsliah, 2008] can be summarized as follows:

1. Let P_G be a set of random vertices in graph G of size $\widetilde{O}(n^{3/4})$, W_G be a set of vertices in graph G of size $\widetilde{O}(n^{1/2})$, and P_H be a set of random vertices in graph H of size $\widetilde{O}(n^{1/4})$.

2. Query all the edges between P_G and W_G in graph G, and all the edges between P_H and V_H in graph H.

3. Enumerate over all the pairs of core sets $C_G \subset P_G$ and $C_H \subset P_H$ such that $|C_G| = |C_H| = (\log n)^2$. For each pair of C_G and C_H,

   (a) Testing identity of the label distribution of vertices of G with respect to C_G and the label distribution of vertices of H with respect to C_H. Since the labels for vertices of H are known, it uses the testing algorithm for one unknown distribution and one unknown distribution [Batu *et al.*, 2013; Paninski, 2008; Valiant and Valiant, 2014], in which $\widetilde{O}(n^{1/2})$ random samples from the unknown distribution are required. Here, it uses the labels of vertices in W_G as the $\widetilde{O}(n^{1/2})$ random samples from the label distribution of vertices in G.

   (b) Testing the second condition of Theorem 6.1.1. The algorithm randomly sample $(\log n)^7$ paris of vertices (u_i, v_i) in G, and randomly find pairs of vertices (u'_i, v'_i) in H satisfying $L_{C_G}(u_i) = L_{C_H}(u'_i)$ and $L_{C_G}(v_i) = L_{C_H}(v'_i)$. It rejects if $e(u_i, v_i) \neq e(u'_i, v'_i)$ for at least $(1 - \varepsilon/2)(\log n)^7$ different i.

4. The algorithm accepts if both 3(a) and 3(b) accepts for at least one pair of C_G and C_H enumerated, otherwise, the algorithm rejects.

Since step 3(a) employs the testing algorithm for one known distribution and one unknown distribution, $|W_G|$ is at least $\Omega(\sqrt{n})$. And the algorithm needs to query $\widetilde{\Theta}(\sqrt{n} \cdot |P_G|)$ edges in graph G and $\widetilde{\Theta}(n \cdot |P_H|)$ edges in graph H. On the other hand, to make sure the existence of the core set pair C_G and C_H satisfying the two conditions of Theorem 6.1.1, it requires that with high probability, there is an isomorphism mapping from vertices of G to vertices of H such that $|\sigma(P_G) \cap P_H| = \Omega(\log n)^2$. Hence, $|P_G| \cdot |P_H| = \Omega(n)$. Hence, the overall sample complexity is optimized by taking $|P_G| = \widetilde{O}(n^{3/4})$ and $|P_H| = \widetilde{O}(n^{1/4})$.

6.1.2 Sketch of our improvement

There are two main bottlenecks to further improving the sample complexity of Fischer and Matsliah's algorithm:

1. The first bottleneck is about testing identity of two label distributions. To make sure there are core sets C_G and C_H satisfying the two conditions of Theorem 6.1.1, we needs

$|P_G| \cdot |P_H| = \Omega(n)$. If we use the algorithm of testing identity of distributions for one known distribution and one unknown distribution, then an $\widetilde{\Omega}(n^{5/4})$ edge query complexity is inevitable. If we use the algorithm of testing identity of distributions for two unknown distributions, then we need to know at least $\Omega(n^{2/3})$ labels in both graphs, according to the $\Omega(n^{2/3})$ sample complexity lower bound by Valiant [Valiant, 2011]. And then we need to query $\Omega(n^{7/6})$ edges in both graphs.

2. The second bottleneck is to randomly sample vertices satisfying the second condition of Theorem 6.1.1. In [Fischer and Matsliah, 2008], the condition is preserved by revealing the labels for all the vertices with respect to the core sets. But this implies an $\widetilde{O}(n^{5/4})$ sample complexity. On the other hand, if we do not query the labels for all the vertices, it is hard to ensure that every pair of vertices in G has the same probability to be sampled, because vertices whose label has high probability in the label distribution are more likely to find a vertex with same label in the other graph.

We get over the two bottlenecks at the same time by making use of the estimation of the *neighbor distance* between every pair of vertices. The neighbor distance between two vertices in the same graph is the normalized distance between the rows of the two vertices in an arbitrary adjacency matrix of the graph (see Section 6.2.1.1 for formal definition). Using a Chernoff bound type argument, one can show that with $\widetilde{O}((n \log n)/\delta^2)$ samples, we can estimate the neighbor distance between any pair of vertices in the graph with additive error at most δ. And the estimation gives us an approximate neighbor distance metric on the vertices of the graph.

Based on the approximate neighbor distance metric, we first choose a parameter r such that for most vertices, the radius r balls (the radius r ball for a vertex is the the set of vertices in the same graph with neighbor distance at most r to the vertex) have a small fraction of vertices on the boundary, and the size of $2r$ radius ball is upper bounded by some subpolynomial multiplicative factor of the size of the r radius ball.

We overcome the first bottleneck by presenting an algorithm to solve the testing label bijection problem. In the testing label bijection problem, we want an algorithm to distinguish the following two cases (see Section 6.5 for the formal statement of the testing label bijection problem):

1. Accept with good probability if there is a bijection from the vertices of G to H such that the labels and the local views of the neighbor distance metric (measured by the mapping distance defined in Section 6.2.1.2) are preserved by the bijection;

2. Reject with good probability if any bijection from the vertices of G to H have a non-negligible fraction of vertices such that either the label or the local views of the neighbor distance metric are not preserved by the bijection.

We reduce the testing label bijection problem to a special instance of testing collision problem between the two graphs. Roughly speaking, a vertex $y \in V_H$ is a collision of vertex $x \in V_G$ if the two vertices have label distance at most r and there exists a vertex $z \in V_H$ with approximate neighbor distance at most r to y such that the local view of the neighbor distance metrics for z is similar to x (see Section 6.4 for the formal definition of collision). In the testing collision problem, given a set S of vertices in graph G, we want an algorithm to accept with good probability if for every vertex x in S, the number of collision to x is close to the size of the r radius ball of x, and all the collisions of x have pairwise neighbor distances upper bounded by about $2r$, and reject with good probability if a non-negligible fraction of vertices do not satisfy the property. Observe that for any positive instance we are going to accept, if $x \in V_G$ has a small fraction of vertices on the boundary of its r radius ball and $y \in V_H$ is a collision of x, the number of collisions to x in the $2r$ radius ball of y is close to the size of the r radius ball of x. Our algorithm makes use of the fact that if $y \in V_H$ is a collision of $x \in V_G$, then using another subpolynomial samples in the $2r$ radius ball of y, one can estimate the number of collisions to x within that ball based on the neighbor distance metric estimation of graph H. This leads to an algorithm with $\widetilde{O}(\sqrt{n})$ label queries in both graphs solving the testing collision problem satisfying that every pair of vertices in S has estimated distance slightly greater than $4r$.

We also carefully sparsify the graph G such that every pair vertices has estimated distance slightly greater than $4r$ after the sparsification. We show that it is sufficient to solve the testing labels problem by solving the testing collision problem on G after sparsification.

In addition, we get rid of the second bottleneck by proposing a new algorithm to sample pairs of vertices $(x \in V_G, y \in V_H)$ such that the label distance between x and y is small, and for most vertices in graph G, the probability of a sampled pair containing the vertex is close

to $1/n$. This new algorithm can be used to near uniformly sample a pair of vertices (u, v) in graph G and another pair of vertices (u', v') in graph H such that the label distances for u, u' and for v, v' are small. Then we are able to test a property similar to the second condition of Theorem 6.1.1.

In the new sampling algorithm, we partition the vertices of graph G into groups such that the vertices within each group have similar sizes of r radius ball. The new sampling algorithm contains two steps. First, it samples a group in graph G with probability proportional to the size of the group. Second, it samples a subset of vertices from the prefixed group in graph G and a set of vertices from graph H. The algorithm returns a pair of sampled vertices $(x \in V_G, y \in V_H)$ such that y is a collision of x.

6.2 Notations and parameters

In this section, we give some definitions and parameters used in the following sections.

6.2.1 Dissimilarity of vertices

An important role in our proof is played by two measures of dissimilarity between vertices. Since each measure defines a metric[1] on vertices of the graph, we simply refer to them as the *distances* between vertices, even though it is unrelated to the standard notion of distance in a graph.

6.2.1.1 Neighbor distance

Let G be an arbitrary unweighted graph. We write V_G and E_G to refer to the set of G's vertices and edges, respectively. For two vertices $v, u \in V_G$, we write $e_G(v, u)$ to denote the indicator whether v and u are connected. Formally,

$$e_G(v,u) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } v \text{ and } u \text{ are connected in } G, \\ 0 & \text{otherwise.} \end{cases}$$

[1]To be more formal, each measure is a semi-metric because the distance between two *different* vertices can be equal to zero.

Note that in particular, for simple graphs, which are considered in this chapter, $e_G(v, v) = 0$, because simple graphs have no self-loops.

For two vertices $v, u \in V_G$, their neighbor distance $d_G(v, u)$ is the normalized number of connections to other vertices they differ at. Formally,

$$d_G(v, u) \stackrel{\text{def}}{=} \frac{|\{w \in V_G : d_G(v, w) \neq d_G(u, w)\}|}{|V_G|}.$$

Intuitively, $d_G(v, w)$ measures how differently two vertices behave with respect to other vertices in the graph.

We describe a query-efficient subroutine for estimating distances between all vertices.

Subroutine `Estimate-Edge-Distances`**:**

Input: graph G, parameter $\sigma \in (0, 1)$

Output: Estimates $\mathcal{M}_G(v, u)$ of $d_G(v, u)$ for all pairs $v, u \in V_G$

1. Select independently uniformly at random m vertices t_1, t_2, ..., t_m in G, where $m \stackrel{\text{def}}{=} \lceil 2 \ln |V_G|/\sigma^2 \rceil$.
2. Query $e_G(u, t_i)$ for all $u \in V_G$ and $i \in [m]$.
3. For every pair $v, u \in V_G$, output

$$\mathcal{M}_G(v, u) \stackrel{\text{def}}{=} \frac{|\{i \in [m] : e_G(v, t_i) \neq e_G(u, t_i)\}|}{m}.$$

Lemma 6.2.1. *Let G be a graph on n vertices. Let $\sigma \in (0, 1)$. With probability $1 - \frac{1}{n^2}$,* `Estimate-Edge-Distances` *computes estimates $\mathcal{M}_G(\cdot, \cdot)$ such that for every pair u and v of vertices,*

$$|d_G(u, v) - \mathcal{M}_G(u, v)| \leq \sigma. \tag{6.1}$$

The query complexity of `Estimate-Edge-Distances` *is $O(n\sigma^{-2} \log n)$.*

Proof. For $u = v$, $d_G(u, v) = \mathcal{M}_G(u, v)$ trivially. Note that both $d_G(\cdot, \cdot)$ and $\mathcal{M}_G(\cdot, \cdot)$ are symmetric, so it suffices that both quantities are within σ for at most $\binom{n}{2}$ pairs. For any pair of vertices u and v, it follows from Hoeffding's inequality that $|d_G(u, v) - \mathcal{M}_G(u, v)| \leq \sigma$

with probability $1 - 2\exp(-2\sigma^2 m) \geq 1 - 2\exp(-4\ln n) = 1 - 2/n^4$. By the union bound, the inequality holds for all u and v with probability at least $1 - \binom{n}{2} \cdot \frac{2}{n^4} \geq 1 - 1/n^2$.

The query complexity of `Estimate-Edge-Distances` is clearly $O(n\sigma^{-2}\log n)$. □

In our algorithm, we run Subroutine `Estimate-Edge-Distances` once for each of the input graphs G and H for some value of σ to be set later. With high probability, the estimated distances between vertices are at distance at most σ from the real values. One can therefore assume that throughout the rest of the proof and throughout the rest of the algorithm's execution, we have nearly correct estimates $\mathcal{M}_G(\cdot,\cdot)$ and $\mathcal{M}_H(\cdot,\cdot)$.

For a graph $I \in \{G, H\}$, a vertex $x \in V_I$, and a radius t, we introduce notation for the ball centered at v of radius t:

$$B_I(x,t) \stackrel{\text{def}}{=} \{v \in V_I : \mathcal{M}_I(x,v) \leq t_1\}.$$

Additionally, for any two radii t_1 and t_2, we introduce notation for the spherical shell between them:

$$B_I(x,t_1,t_2) \stackrel{\text{def}}{=} \{v \in V_I : t_1 < \mathcal{M}_I(x,v) \leq t_2\}.$$

6.2.1.2 Map distance

Let G and H be two arbitrary graphs with same number of vertices. We define the map distance between two vertices in $V_G \cup V_H$(not necessarily in same graph) as following:

Definition 6.2.2. *Let $I, J \in \{G, H\}$. Given $\mathcal{M}_I$,$\mathcal{M}_J$ and two vertices $x \in V_I, y \in V_J$. Denote Π be the set of all the bijections from V_I to V_J. Let*

$$\rho_0(x,y) = \min_{\pi \in \Pi : \pi(x) = y} \max_{u \in V_I} |\mathcal{M}_I(x,u) - \mathcal{M}_J(y,\pi(u))|,$$

and for $i \geq 1$,

$$\rho_i(x,y) = \min_{\pi \in \Pi : \pi(x) = y} \max_{u \in V_I} \{|\mathcal{M}_I(x,u) - \mathcal{M}_J(y,\pi(u))|, |\rho_{i-1}(u) - \rho_{i-1}(\pi(u))|\}.$$

Subroutine `Estimate-Map-Distance`**:**

Input: Two graphs G, H, $\mathcal{M}_G, \mathcal{M}_H$, parameter i and σ_0

Output: A function $\gamma_{i,\sigma_0} : \{V_G \cup V_H\}^2 \to \{0, 1\}$.

1. Let x, y be two vertices in $V_G \cup V_H$, $I \in \{G, H\}$ be the graph containing x, and $J \in \{G, H\}$ be the graph containing y. For any x, y, build a bipartite graph $B_{0,x,y}$ on (V_I, V_J) such that two vertices $u \in V_I, v \in V_J$ are adjacent iff $|\mathcal{M}_I(x, u) - \mathcal{M}_J(y, v)| \leq \sigma_0$. Let $\gamma_{0,\sigma_0}(x, y) = 1$ if there is a perfect matching on $B_{0,x,y}$ with x matching y, otherwise $\gamma_{0,\sigma_0}(x, y) = 0$.

2. For any $1 \leq j \leq i$, and any two vertices x, y, build a bipartite graph $B_{j,x,y}$ on (V_I, V_J) such that two vertices $u \in V_I, v \in V_J$ are adjacent iff

$$|\mathcal{M}_I(x, u) - \mathcal{M}_J(y, v)| \leq \sigma_0 \text{ and } \gamma_{j-1,\sigma_0}(u, v) = 1.$$

   Let $\gamma_{j,\sigma_0}(x, y) = 1$ iff there is a perfect matching on $B_{j,x,y}$ with x matching y,, otherwise $\gamma_{j,\sigma_0}(x, y) = 0$.

By the definition of map distance, we have

Fact 6.2.3. *$\gamma_{i,\sigma_0}(x, y) = 1$ iff $\rho_i(x, y) \leq \sigma_0$, otherwise $\gamma_{i,\sigma_0}(x, y) = 0$.*

6.2.2 Parameters

Given ε_0, let

$$\varepsilon = \varepsilon_0/1000, \phi = \exp\left(-\sqrt{\log n}\right),$$

$$\varepsilon_1 = \frac{1}{\log n}, \varepsilon_2 = \frac{1}{\log^{100} n}, \sigma = \exp\left(-\frac{100\sqrt{\log n}\log\log n}{\varepsilon}\right), \mu = \frac{128\sigma\log^4 n}{\varepsilon_1}.$$

Given a parameter r, we define

$$Z_I = \{v : |B_I(v, r + 4\mu)|/|B_I(v, r - 4\mu)| \leq 1 + \varepsilon_1/4$$
$$\text{and } |B_I(x, 1600r\log n/\varepsilon_1 + 4\mu)|/|B_I(x, r - 4\mu)| \leq 1/4\phi\},$$

$$A_I = \{v : |B_I(v, r + 3\mu)|/|B_I(v, r - 3\mu)| \leq 1 + \varepsilon_1/3$$
$$\text{and } |B_I(x, 1600r\log n/\varepsilon_1 + 3\mu)|/|B_I(x, r - 3\mu)| \leq 1/3\phi\},$$

$$S_I = \{v : |B_I(v, r + 2\mu)|/|B_I(v, r - 2\mu)| \leq 1 + \varepsilon_1/2$$
$$\text{and } |B_I(x, 1600r\log n/\varepsilon_1 + 2\mu)|/|B_I(x, r - 2\mu)| \leq 1/2\phi\},$$

$$U_I = \{v : |B_I(v, r+\mu)|/|B_I(v, r-\mu)| \leq 1+\varepsilon_1 \\ \text{and } |B(v, 1600r\log n/\varepsilon_1 + \mu)|/|B(v, r-\mu)| \leq 1/\phi\}.$$

We select the parameter $r < \phi^{100}$ such that both $|Z_G|$ and $|Z_H|$ are at least $(1-\varepsilon)n$.

Lemma 6.2.4. *Given two graphs G, H and corresponding approximate distance metrics $\mathcal{M}_I$ and $\mathcal{M}_J$, there exists a parameter $\frac{1000\mu\log n}{\varepsilon_1^2} < r < \phi^{100}$ such that $|Z_G| \geq (1-\varepsilon)n$ and $|Z_H| \geq (1-\varepsilon)n$.*

Proof. Let $\alpha = \frac{3200\log n}{\varepsilon_1}$, $d_0 = \frac{1000\mu\log n}{\varepsilon_1^2}$ and $d_{i+1} = \alpha d_i$ for $0 \leq i \leq m$ with $m = \lceil \frac{12\sqrt{\log n}}{\varepsilon} \rceil$. We have $d_{m+1} < \phi^{100}$. For every vertex $x \in V_G$, there are at most $\lceil \log_{1/4\phi} n \rceil < 2\sqrt{\log n}$ different i satisfying $B_G(x, d_{i+1})/B_G(x, d_i) > 1/4\phi$. Let $G_i = \{x \in G : B_G(x, d_{i+1})/B_G(x, d_i) \leq 1/4\phi\}$. There are at most $\frac{2\sqrt{\log n}n}{\varepsilon n/2} = \frac{4\sqrt{\log n}}{\varepsilon}$ different i such that $|G_i| < (1-\frac{\varepsilon}{2})n$. Hence, at least $m - \frac{4\sqrt{\log n}}{\varepsilon} \geq \frac{2m}{3}$ different i satisfying $|G_i| \geq (1-\frac{\varepsilon}{2})n$.

Similarly, let $H_i = \{x \in H : B_H(y, d_{i+1})/B_H(y, d_i) \leq 1/4\phi\}$. There are at least $2m/3$ different i such that $|H_i| \geq (1-\frac{\varepsilon}{2})n$. So, there exists an i such that both $|G_i| \geq (1-\frac{\varepsilon}{2})n$ and $|H_i| \geq (1-\frac{\varepsilon}{2})n$ hold. We arbitrarily fix such an i.

Now let $c_j = d_i + 8\mu j + 4\mu$ for $0 \leq j \leq m'$ where $m' = \lceil \frac{30\log n}{\varepsilon_1\varepsilon} \rceil$, and let

$$G_{i,j} = \{x \in G_i : B_G(x, c_j + 4\mu)/B_G(x, c_j - 4\mu) \leq 1 + \varepsilon_1/4\}.$$

Every vertex in G_i does not belong to at most $5\log n/\varepsilon_1$ different $G_{i,j}$. There are at most $\frac{10\log n}{\varepsilon_1\varepsilon}$ different i satisfying $|G_{i,j}| < |G_i| - \frac{\varepsilon n}{2}$. There exists at least $m' - \frac{10\log n}{\varepsilon_1\varepsilon} \geq \frac{2m'}{3}$ different j such that $|G_{i,j}| \geq |G_i| - \frac{\varepsilon n}{2}$.

Similarly, there are at least $2m'/3$ different j such that $H_{i,j} \geq H_i - \frac{\varepsilon n}{2}$, where $H_{i,j} = \{y \in H_i : B_H(y, c_j + 2\mu)/B_H(x, c_j - 2\mu) \leq 1 + \varepsilon_1/4\}$.

So, there exists a j such that both $|G_{i,j}| \geq (1-\varepsilon)n$ and $|H_{i,j}| \geq (1-\varepsilon)n$ hold. Let $r = c_j$. The lemma follows. □

6.2.3 Weight functions

Given two graphs G and H with approximate distance metrics $\mathcal{M}_G$ and $\mathcal{M}_H$. We consider function $w : (V_G \times V_G) \cup (V_H \times V_H) \to \mathbb{R}^{\geq 0}$. We say w is a weight function for G and H if $\sum_{v,x \in V_G} w(v,x) \leq n$ and $\sum_{y,z \in V_H} w(y,z) \leq n$. We always denote $w(x) = \sum_{v \in V_G} w(x,v)$ for

any $x \in V_G$ and $w(y) = \sum_{z \in V_H} w(y,z)$ for any $z \in V_H$. We say a weight function is robust if for any $u, v \in V_G \cup V_H$, $\rho_2(u,v) \leq 4\delta$ implies that $(1-\varepsilon_1)w(v) \leq w(u) \leq (1+\varepsilon_1)w(v)$.

6.3 Sparsification

In this section, we present an algorithm to *sparsify* vertices in a graph such that after the sparsification, every pair of remaining vertices has neighbor distance estimation at least $4r + 6\delta$.

Given a robust weight function w on graph I and J.

Subroutine `Sparsification`

Input: Two graph I, J, metrics $\mathcal{M}_I, \mathcal{M}_J$, a set of vertex $S_I \subseteq V_I$, weight functions $w : V_I \cup V_J \to \mathbb{R}^{\geq 0}$ such that $\sum_{x \in V_I} w(x) \leq n$ and $\sum_{y \in V_J} w(y) \leq n$.

Output: Accept or reject. If accept, then also return sets $S_{i,j,k}, T_{i,j,k} \subseteq V_I$, $H_{i,j,k}, C_{i,j,k} \subseteq V_J$ for every $0 \leq i,j,k \leq 6\log n/\varepsilon_1$, and $\mathsf{weight}_w : \cup_{i,j,k} T_{i,j,k} \to \mathbb{R}^{\geq 0}$.

1. Partition S_I into $S_{i,k}$ for $0 \leq i,k \leq 6\log n/\varepsilon_1$, where $S_{i,k} = \{v \in S_I : (1+\varepsilon_1)^i/n^2 \leq w(v) < (1+\varepsilon_1)^{i+1}/n^2, (1+\varepsilon_1)^k \leq B_I(v,r) < (1+\varepsilon_1)^{k+1}\}$
2. For every $0 \leq i,k \leq 6\log n/\varepsilon_1$,
   (a) Let $T_{i,k} = \emptyset$ initially.
   (b) Repeat following process for an arbitrary order of vertices in $S_{i,k}$: for every vertex $x \in S_{i,k}$, if the $\mathcal{M}_I$ distance between any vertex in $T_{i,k}$ and x is more than $4r + 6\sigma$, then add x in $T_{i,k}$.
   (c) For every vertex $x \in S_{i,k}$, let $Assignto(x)$ be the vertex in $T_{i,k}$ with smallest distance to x in $\mathcal{M}_I$(if there are more than one vertex with smallest distance, use an arbitrary one). For a vertex $v \in T_{i,k}$, let $Assign(v) = \{x \in S_{i,k} : v = Assignto(x)\}$.
   (d) Partition $T_{i,k}$ into $T_{i,j,k}$ for $0 \leq j \leq 6\log n/\varepsilon_1$ such that $v \in T_{i,k}$ is in $T_{i,j,k}$ if $(1+\varepsilon_1)^j \leq |Assign(v)| < (1+\varepsilon_1)^{j+1}$.

3. For every $0 \le i,j,k \le 6\log n/\varepsilon_1$, let $\alpha_{i,j,k} = (1+\varepsilon_1)^i/n^2$, $\beta_{i,j,k} = (1+\varepsilon_1)^j$, $\gamma_{i,j,k} = (1+\varepsilon_1)^k$,

$$H_{i,j,k} = \{y \in V_J : (1-6\varepsilon_1)\alpha_{i,j,k} \le w(y) \le (1+6\varepsilon_1)(1+\varepsilon_1)\alpha_{i,j,k},$$
$$(1-\varepsilon_1)\gamma_{i,j,k} \le |B_J(y,r)| \le (1+\varepsilon_1)^2\gamma_{i,j,k}\}$$

$$C_{i,j,k} = \{z \in V_J : \exists x \in T_{i,j,k}, y \in H_{i,j,k} \text{ s.t. } \rho_2(x,y) \le 2\delta \text{ and } \mathcal{M}_J(y,z) \le 50r\}.$$

Reject if $|C_{i,j,k}| < |T_{i,j,k}|\gamma_{i,j,k}$.

4. For every $0 \le i,j,k \le 6\log n/\varepsilon_1$ and any $x \in T_{i,j,k}$, let $\mathsf{weight}_w(x) = \sum_{v \in Assign(x)} w(v)$.

5. Accept.

Denote $\mathsf{weight}_w(T_{i,j,k}) = \sum_{v \in T_{i,j,k}} \mathsf{weight}_w(x)$ for any $0 \le i,j,k \le 6\log n\varepsilon_1$.

Lemma 6.3.1. *If subroutine* `Scarification` *rejects, then there does not exists a bijection $\pi : V_I \to V_J$ satisfying for any $w,x \in V_I$, $\mathcal{M}_J(\pi(w),\pi(x)) - 2\sigma \le \mathcal{M}_I(w,x) \le \mathcal{M}_J(\pi(w),\pi(x)) + 2\sigma$.*

Proof. We prove by contradiction. Assume there is a π as we want. For any $x \in V_I$, $\rho_2(x,\pi(x)) \le 2\sigma$, and thus $(1-\varepsilon_1)w(\pi(x)) \le w(x) \le (1+\varepsilon_1)w(\pi(x))$.

Hence, for any $T_{i,j,k}$, $\pi(T_{i,j,k}) \subseteq H_{i,j,k}$. For every pair of vertices $y,z \in T_{i,j,k}$,

$$\mathcal{M}_J(\pi(y),\pi(z)) \ge \mathcal{M}_I(y,z) - 2\sigma > 4r + 4\sigma.$$

On the other hand, since for any $x \in T_{i,j,k}$ $B_I(x,r) \ge \gamma_{i,j,k}$ and $x \in S_I$,

$$|C_{i,j,k}| \ge \sum_{y \in \pi(T_{i,j,k})} |B_J(y, r+2\delta)| \ge \sum_{x \in T_{i,j,k}} |B_I(x,r)| \ge |T_{i,j,k}|\gamma_{i,j,k}$$

□

Fact 6.3.2. *For any vertex $x \in T_{i,k}$, $(B_I(x, 2r+2\sigma) \cap S_{i,k}) \subseteq Assign(x)$.*

Lemma 6.3.3. $|H_{i,j,k}| \le \min\{n, \frac{n}{(1-6\varepsilon_1)\alpha_{i,k}}\}$ *and* $|C_{i,j,k}| \le \min\{n, \frac{n\gamma_{i,j,k}(1+\varepsilon_1)^2}{(1-6\varepsilon_1)\alpha_{i,j,k}\beta_{i,j,k}\phi}\}$.

Proof. Since every $y \in H_{i,j,k}$ satisfying $w(y) \geq (1-6\varepsilon_1)\alpha_{i,j,k}$ and $\sum_{y\in V_J} w(y) \leq n$, we have $|H_{i,j,k}| \leq \min\{n, \frac{n}{(1-6\varepsilon_1)\alpha_{i,k}}\}$.

To prove the upper bound of $|C_{i,j,k}|$, we first find a sequence of vertices in $H_{i,j,k}$, denoted as $h_1, h_2, ...h_m$, satisfying

1. For every h_ℓ, there exists a $x \in T_{i,j,k}$ such that $\rho_2(x, h_\ell) \leq 4\delta$.

2. $\mathcal{M}_J(h_\ell, h_t) \geq 9r$ for $\ell \neq t$.

Since for every vertex $x \in T_{i,j,k}$ there are at least $\beta_{i,j,k}$ vertices in $S_{i,j,k}$ with distance at most $4r + 6\sigma$ to x in V_I, there are also at least $\beta_{i,j,k}$ vertices in $H_{i,j,k}$ with distance at most $4r + 10\sigma$ to h_ℓ in V_J. Hence $m \leq |H_{i,j,k}|/\beta_{i,j,k} \leq \frac{n}{(1-6\varepsilon_1)\alpha_{i,j,k}\beta_{i,j,k}}$. So

$$|C_{i,j,k}| \leq |\cup_{\ell=1}^{m} B_J(h_\ell, 50r)| \leq \min\{n, \frac{n\gamma_{i,j,k}(1+\varepsilon_1)^2}{(1-6\varepsilon_1)\alpha_{i,j,k}\beta_{i,j,k}\phi}\}$$

□

Lemma 6.3.4. *If* $\mathsf{weight}_w(T_{i,j,k}) \geq \varepsilon_2^2 n$, *then*

$$\frac{\varepsilon_2^2 n}{2\alpha_{i,j,k}\beta_{i,j,k}} \leq |T_{i,j,k}| \leq \frac{n}{\alpha_{i,j,k}\beta_{i,j,k}}$$

Proof. For every vertex $x \in T_{i,j,k}$, since $\mathsf{weight}_w(x) = \sum_{v\in Assign(v)} w(v)$ and $\beta_{i,j,k} \leq |Assign(v)| \leq (1+\varepsilon_1)\beta_{i,j,k}$, we have

$$\alpha_{i,j,k}\beta_{i,j,k} \leq \mathsf{weight}_w(x) \leq (1+\varepsilon_1)^2\alpha_{i,j,k}\beta_{i,j,k}.$$

Hence

$$\frac{\varepsilon_2^2 n}{2\alpha_{i,j,k}\beta_{i,j,k}} \leq \frac{\varepsilon_2^2 n}{(1+\varepsilon_1)^2\alpha_{i,j,k}\beta_{i,j,k}} \leq |T_{i,j,k}| \leq \frac{n}{\alpha_{i,j,k}\beta_{i,j,k}}$$

□

By Lemma 6.3.3 and 6.3.4, we have

Corollary 6.3.5. *Assume* $\mathsf{weight}_w(T_{i,j,k}) \geq \varepsilon_2^2 n$. *We have*

$$|T_{i,j,k}|\gamma_{i,j,k} \leq |C_{i,j,k}| \leq 4|T_{i,j,k}|\gamma_{i,j,k}/\varepsilon_2^2\phi.$$

If $|T_{i,j,k}| \leq \sqrt{n}/\phi^{14}$, *then* $\gamma_{i,j,k} \geq \varepsilon_2^2\phi^{15}|C_{i,j,k}|/4\sqrt{n}$. *If* $|T_{i,j,k}| > \sqrt{n}/\phi^{14}$, *then* $\gamma_{i,j,k} < 2\phi^{14}|C_{i,j,k}|/\sqrt{n}$.

Lemma 6.3.6. *If* $\mathsf{weight}_w(T_{i,j,k}) \geq \varepsilon_2^2 n$, *then* $\gamma_{i,j,k} \leq 2\alpha_{i,j,k}\beta_{i,j,k}/\varepsilon_2^2$.

Proof. Since every pair of vertices in $T_{i,j,k}$ has distance at least $4r + 6\sigma$, and every vertex $x \in T_{i,j,k}$ satisfies $|B_I(x, r)| \geq \gamma_{i,j,k}$, by Lemma 6.3.4,

$$\frac{\varepsilon_2^2 n}{2\alpha_{i,j,k}\beta_{i,j,k}}\gamma_{i,j,k} \leq |\cup_{x\in T_{i,j,k}} B_I(x, r)| \leq n$$

□

6.4 Testing collision

In this and the next section, we assume that every vertex in graphs G and H is associated with a fixed binary string of length, called label. The length of these binary strings will be specified later. For any pair of vertices $x, y \in V_I \cup V_J$, let the label distance between the two vertices, denoted as $\mathcal{M}(x, y)$, be the hamming distance of the labels of the two vertices.

In this section, we present an algorithm for the testing collision problem. The main reason of studying the testing collision problem is that the problem of testing label bijection can be reduced to the testing collision problem, and the problem of testing label bijection will be used to bypass the $\Omega(n^{2/3})$ lower bound of testing identity of two unknown distributions using the estimation of neighbor distance metric.

We start from the definition of collision.

Definition 6.4.1. *Let* $I, J \in \{G, H\}$ *be two different graphs. A vertex* $y \in V_J$ *is a collision to* $x \in V_I$ *if*

1. $\mathcal{M}(x, y) \leq r$
2. *There is a vertex* $z \in V_J$ *satisfying* $\mathcal{M}_J(y, z) \leq r + 2\sigma$ *and* $\rho_4(x, z) \leq 2\sigma$.

A vertex $y \in V_J$ *is a good collision with vertex* $x \in V_I$ *if*

1. y *is a collision with* x;
2. *Every vertex* $z \in B_J(y, 2r + 6\sigma)$ *satisfying* $\mathcal{M}(x, z) \leq r$ *is a collision of* x;
3. $(1 - \varepsilon_1)|B_I(x, r)| \leq |N(x, y, 2r + 6\sigma)| \leq (1 + \varepsilon_1)|B_I(x, r)|$, *where* $N(x, y, \alpha) = \{z \in B_J(y, \alpha) : z$ *is a collision with* $x\}$;

4. *There is no* $z \in B_J(y, 2r + 6\sigma, 29r)$ *satisfies* $\mathcal{M}(x, z) \leq r$.

A vertex $y \in V_J$ *is a bad collision with vertex* $x \in V_I$ *if* y *is a collision with* x *and at least one of following conditions hold*

1. $|N(x, y, 2r + 6\delta)| < (1 - 2\varepsilon_1)|B_I(x, r)|$ *or* $|N(x, y, 2r + 6\delta)| > (1 + 2\varepsilon_1)|B_I(x, r)|$*;*

2. *At least* $\varepsilon_1|B_I(x, r)|$ *vertices in* $B_J(y, 2r + 6\delta)$ *have distance at most* r *to* x *in* $\mathcal{M}$*, but not collisions of* x.

3. *At least* $\varepsilon_1|B_I(x, r)|$ *vertices in* $B_J(y, 2r + 6\delta, 29r)$ *have distance at most* r *to* x *in* $\mathcal{M}$.

A vertex $y \in V_J$ *is an intermediate collision with vertex* $x \in V_I$ *if* y *is a collision with* x*, but neither good nor bad.*

Fact 6.4.2. *If* $y \in C_{i,j,k}$ *is a good or intermediate collision to* $v \in T_{i,j,k}$*, then there are at least* $(1 - 2\varepsilon_1)|B_I(v, r)|$*(or* $(1 - 2\varepsilon_1)\gamma_{i,j,k}$*) and at most* $(1 + 2\varepsilon_1)|B_I(v, r)|$*(or* $(1 + 2\varepsilon_1)(1 + \varepsilon_1)\gamma_{i,j,k}$*) collisions to* v *in* $B_J(y, 2r + 6\sigma)$*, and then there are at most* $(1 + 3\varepsilon_1)|B_I(v, r)|$*(or* $(1 + 3\varepsilon_1)(1 + \varepsilon_1)\gamma_{i,j,k}$*) collisions to* v *in* $B_J(y, 29r)$.

Definition 6.4.3. *A vertex* $y \in V_J$ *is a useful collision to* $x \in V_I$ *if*

1. y *is a good collision to* x.

2. *For every* $z \in B_J(y, 2r + 6\sigma)$ *such that* z *is a collision to* x*,* z *is a good collision to* x.

y *is a useless collision to* x *if one of following conditions hold*

1. y *is a bad collision to* x.

2. *At least* $\varepsilon_1|B_I(v, r)|$ *vertices in* $B_J(y, 2r + 6\sigma)$ *are bad collisions to* x.

A vertex y *is a semi-useful collision to* x *if it is not a useless collision to* x.

Fact 6.4.4. *If* $y \in C_{i,j,k}$ *is a useful collision to* $v \in T_{i,j,k}$*, then there are at least* $(1 - \varepsilon_1)|B_I(v, r)|$*(or* $(1-\varepsilon_1)\gamma_{i,j,k}$*) and at most* $(1+\varepsilon_1)|B_I(v, r)|$*(or* $(1+\varepsilon_1)^2\gamma_{i,j,k}$*) good collisions to* v *in* $B_J(y, 2r + 6\sigma)$.

If $y \in C_{i,j,k}$ is a semi-useful collision to $v \in T_{i,j,k}$, then there are at least $(1-3\varepsilon_1)|B_I(v,r)|$ (or $(1-3\varepsilon_1)\gamma_{i,j,k}$) and at most $(1+2\varepsilon_1)|B_I(v,r)|$ (or $(1+2\varepsilon_1)(1+\varepsilon_1)\gamma_{i,j,k}$) good/intermediate collisions to v in $B_J(y, 2r+6\sigma)$.

If $y \in C_{i,j,k}$ is a semi-useful collision to $v \in T_{i,j,k}$, then there are at most $(1+3\varepsilon_1)|B_I(v,r)|$(or $(1+3\varepsilon_1)(1+\varepsilon_1)\gamma_{i,j,k}$) good/intermediate collisions to v in $B_J(y, 29r)$.

In the following of this section, we present an algorithm to solve the following problem.

Testing-Collision Problem

Input: Two graphs I and J, vertex subsets $S_{i,j,k}, T_{i,j,k} \subseteq V_I, C_{i,j,k}, H_{i,j,k} \subseteq V_J$ for $0 \leq i,j,k \leq 6\log n/\varepsilon_1$, $\mathsf{weight}_w : \cup T_{i,j,k} \to \mathbb{R}^{\geq 0}$ satisfying $\sum_{x \in \cup T_{i,j,k}} \geq (1-\varepsilon)n$, label query oracle $\mathcal{O}$, parameter δ.

Output:

1. Accept with probability $1-\delta$ if all of the following conditions hold
   (a) For both I and J, every pair of vertices from same graph has distance not distorted.
   (b) For every $0 \leq i,j,k \leq 6\log n/\varepsilon_1$, and $x \in T_{i,j,k}, y \in C_{i,j,k}$ satisfying $\mathcal{M}(x,y) \leq r$, y is a useful collision of x.
   (c) Every vertex $x \in \cup T_{i,j,k}$, x has at least $(1-\varepsilon_1)B_I(x,r)$ and at most $(1+\varepsilon_1)B_I(x,r)$ useful collisions.
2. Reject with probability $1-\delta$ if a total weight of at least $12\varepsilon_1 \sum_{i,j,k} \mathsf{weight}_w(T_{i,j,k})$ vertices in $\cup T_{i,j,k}$ do not have semi-useful collision.

In the following of this section, we prove the following theorem.

Theorem 6.4.5. *If for every $T_{i,j,k}$, every pair of vertices has neighbor distance estimation at least $4r+6\delta$, then there is an algorithm solving the Testing Collision-Problem with probability at least $1-\delta$ using $O((\sqrt{n}\log n)\cdot\frac{1}{\phi^{O(1)}}\cdot\log(1/\delta)\cdot\frac{1}{\varepsilon_2^2})$ label queries with running time poly$(n, \frac{1}{\phi}, \frac{1}{\varepsilon_2^2}, \log(1/\delta))$.*

6.4.1 Testing distance distortion

We first present two useful subroutines to check whether the label distances between most pairs of vertices in the same graph are close to their estimated distances.

6.4.1.1 Testing vertex distortion

Definition 6.4.6. *Let $I \in \{G, H\}$. The distance between two vertices $v, u \in V_I$ is distorted by f_I if $\mathcal{M}_I(v,u) < \mathcal{M}(v,u) - 2\sigma$ or $\mathcal{M}_I(v,u) > \mathcal{M}(v,u) + 2\sigma$.*

Definition 6.4.7. *Let $I \in \{G, H\}$ be a graph, and x be a vertex in V_I. For $0 \leq i \leq 6\log n/\varepsilon_1$, denote*

$$\Psi_{x,i} = \{v \in V_I : (1+\varepsilon_1)^i/n^2 \leq \mathcal{M}_I(x,v) < (1+\varepsilon_1)^{i+1}/n^2\}$$

,and

$$\Lambda_{x,i} = \{v \in \Psi_{x,i} : \mathcal{M}(x,v) - 2\sigma \leq \mathcal{M}_I(x,u) \leq \mathcal{M}(x,v) + 2\sigma\}.$$

We say x is λ-distorted by f_I if

1. *For all the $s \in \{\mathbf{uni}, \mathbf{fi}, \bar{\mathbf{fi}}\}$, and every $0 \leq i \leq 6\log n/\varepsilon_1$ with $\sum_{v\in\Psi_{x,i}} s(x,v) > 0$, $\frac{\sum_{v\in\Lambda_{x,i}} s(x,v)}{\sum_{v\in\Psi_{x,i}} s(x,v)} \geq 1-\lambda$.*
2. *There is a $s \in \{\mathbf{uni}, \mathbf{fi}, \bar{\mathbf{fi}}\}$ and an $0 \leq i \leq 6\log n/\varepsilon_1$ with $\sum_{v\in\Psi_{x,i}} s(x,v) > 0$ such that $\frac{\sum_{v\in\Lambda_{x,i}} s(x,v)}{\sum_{v\in\Psi_{x,i}} s(x,v)} = 1-\lambda$.*

Subroutine `Testing-Single-Vertex-Distortion`**:**

Input: A graph I, a vertex $x \in V_I$, weight functions $\mathbf{uni}, \mathbf{fi}, \bar{\mathbf{fi}} : V_I \times V_I \to R^{\geq 0}$ and parameters λ, δ.

Output: Accept or reject.

1. For every $s \in \{\mathbf{uni}, \mathbf{fi}, \bar{\mathbf{fi}}\}$ and every $0 \leq i \leq 6\log n/\varepsilon_1$, randomly sample $\frac{2\log(1/\delta)}{\lambda}$ vertices v in $\Psi_{x,i}$ with probability proportional to $s(x,v)$. Reject if $\mathcal{M}(x,v) \geq \mathcal{M}_I(x,v) - 2\sigma$ or $\mathcal{M}(x,v) \leq \mathcal{M}_I(x,v) + 2\sigma$.
2. Accept.

Lemma 6.4.8. *Let x be a vertex in I.*

1. *If for every v, $\mathcal{M}(x,v)-2\sigma \leq \mathcal{M}_I(x,u) \leq \mathcal{M}(x,v)+2\sigma$, then* `Testing-Single-Vertex-Distortion` *accept with probability 1.*

2. *If a vertex x is at least λ-distorted, then with probability at least $1-\delta$,* `Testing-Single-Vertex-Distortion` *rejects.*

Proof. The first case is obvious. For the second case, since x is at least λ-distorted, then there is a $s \in \{\mathbf{uni}, \mathbf{fi}, \bar{\mathbf{fi}}\}$ and a $0 \leq i \leq 6\log n/\varepsilon_1$ such that $\frac{\sum_{v\in\Lambda_{x,i}} s(x,v)}{\sum_{v\in\Psi_{x,i}} s(x,v)} < 1-\lambda$. The probability of sampling a vertex in $\Lambda_{x,i} - \Psi_{x,i}$ is at least λ. Hence, the probability that none of the sampled vertices are in $\Lambda_{x,i} - \Psi_{x,i}$ is at most

$$(1-\lambda)^{2\log(1/\delta)/\lambda} \leq \delta.$$

□

Subroutine `Testing-Vertex-Distortion`**:**

Input: A graph I, a set of vertices $S \subseteq V_I$, weight function $w : V_I \times V_I \to \mathbb{R}^{\geq 0}$(assuming $w(x) = \sum_{v\in V_I} w(x,v)$ for any $x \in V_I$), and parameters λ, δ, α.

1. Randomly sample a set of $8\sqrt{n}\log(1/\delta)/\alpha$ vertices. The probability of sample $x \in S$ is $\frac{w(x)}{\sum_{v\in S} w(v)}$. Run Subroutine `Testing-Single-Vertex-Distortion` with each sampled vertex with weight function w and parameters $\lambda, \delta/2$. Reject if any execution of Subroutine `Testing-Single-Vertex-Distortion` rejects.

2. Accept.

Lemma 6.4.9. *If every pair of vertices $v, u \in S$ satisfying $\mathcal{M}_I(v,u) - 2\sigma \leq \mathcal{M}(v,u) \leq \mathcal{M}_I(v,u) + 2\sigma$, then Subroutine* `Testing-Vertex-Distortion` *accepts with probability* 1.

If a total weight of at least $\frac{\alpha\sum_{x\in S} w(x)}{\sqrt{n}}$ vertices are at least λ-distorted, then Subroutine `Testing-Vertex-Distortion` *rejects with probability at least $1 - \exp(-1/\phi^4)$.*

Proof. The first case is easy. Now, we prove the second case. If a total weight of at least $\frac{\alpha \sum_{x \in S} w(x)}{\sqrt{n}}$ vertices are at least λ-distorted, then the probability that none of the sampled vertices is at least λ-distorted is at most

$$(1 - \frac{\alpha}{\sqrt{n}})^{8\sqrt{n}\log(1/\delta)/\alpha} \leq \delta/2.$$

By Lemma 6.4.8, if there is one sampled vertex at least λ-distorted, then with probability $1 - \delta/2$, Subroutine `Testing-Single-Vertex-Distortion` rejects. The lemma is obtained by union bound. □

6.4.1.2 Testing set distortion

Given a set of vertices S in graph I explicitly.

Definition 6.4.10. *A set $M \subseteq S$ is a distorted set of S with respect to f_I if at least one of the following conditions hold*

1. *$|M| \geq \phi^{16}|S|/\sqrt{n}$ vertices such that for every $u \in M$, at least $1/100$ fraction of all the vertices v in M satisfy $\mathcal{M}_I(u,v) > \mathcal{M}(u,v) + 2\sigma$ or $\mathcal{M}_I(u,v) < \mathcal{M}(u,v) - 2\sigma$.*

2. *$|M| \geq 2|S|/\phi\sqrt{n}$ vertices in S such that for every vertex $u \in M$, there are at least $\frac{2\phi^3|S|^2}{|M|n}$ and at most $\phi^4|S|/\sqrt{n}$ vertices v in M satisfying*

$$\mathcal{M}_I(u,v) > \mathcal{M}(u,v) + 2\sigma \text{ or } \mathcal{M}_I(u,v) < \mathcal{M}(u,v) - 2\sigma, \tag{6.2}$$

3. *There are $A \subseteq S$ and $B \subseteq V_I$ such that $|A| \geq \max\{1, |S|/2\sqrt{n}\}$, $|B| \geq \sqrt{n}/4$ and for every $x \in A, y \in B$, the distance between x and y is distorted.*

4. *There is a function $s : S \times V_I \to \mathbb{R}^{\geq 0}$ such that*

   (a) *If $s(x,y) > 0$, then the distance between x and y is distorted.*

   (b) *for any $x \in S$, $\sum_{y \in V_I} s(x,y) \leq 2n/|S|$;*

   (c) *for any $y \in V_I$, $\sum_{x \in S} s(x,y) \leq \sqrt{n}$;*

   (d) *$\sum_{x,y} s(x,y) \geq \varepsilon_1 \phi n/10 \log n$.*

Subroutine `Testing-Distance-Preserved-Subset`**:**

Input: A set of vertices $S \subseteq V_I$, parameter δ

1. Randomly sample a set of $400\log(1/\delta)\sqrt{n}/\phi^{16}$ vertices in S, and $400\log(1/\delta)\sqrt{n}/\phi^{16}$ vertices in V_I reject if there exists two sampled vertices $u \in S, v \in V_I$ such that $\mathcal{M}(u,v) < \mathcal{M}_I(u,v) - 2\sigma$ or $\mathcal{M}(u,v) > \mathcal{M}_I(u,v) + 2\sigma$.
2. Accept.

Lemma 6.4.11. *If for every pair of vertices $u \in S$ and $v \in V_I$, the distance between u and v is not distorted, then Subroutine* `Testing-Distance-Preserved-Subset` *accepts with probability* 1.

If S has a distorted set M, then Subroutine `Testing-Distance-Preserved-Subset` *rejects with probability $1-\delta$.*

Proof. It is straightforward that if the distance between every pair of vertices $u \in S$ and $v \in V_I$ is not distorted, then Subroutine `Testing-Distance-Preserved-Subset` always accepts.

We first show that if the first condition of Definition 6.4.10 holds for M, then the subroutine rejects with probability at least $1-\delta$. The probability that one sample is in M is at least $\phi^{16}/\sqrt{n}$. So, the probability that the first half of all the samples contain at least one vertex $u \in M$ is at least

$$1 - \left(1 - \frac{\phi^{16}}{\sqrt{n}}\right)^{400\log(1/\delta)\sqrt{n}/\phi^{16}} \geq 1 - \delta/2.$$

We bound the probability that the second half of all the samples contain one v satisfying $\mathcal{M}_I(u,v) > \mathcal{M}(u,v) + 2\sigma$ or $\mathcal{M}_I(u,v) < \mathcal{M}(u,v) - 2\sigma$. The probability of such a v is at least $\frac{|M|}{100|S|} \geq \frac{\phi^{16}}{100\sqrt{n}}$. So, the probability that second half of all the samples do not contain such an v is at most

$$(1 - \frac{\phi^{16}}{100\sqrt{n}})^{400\log(1/\delta)\sqrt{n}/\phi^{16}} \leq \delta/2.$$

The lemma holds by union bound.

Now we show that if the second condition of Definition 6.4.10 holds for M, then the subroutine rejects with probability at least $1-\delta$. Consider the probability that a sequence of $\sqrt{n}/\phi^3$ samples find a pair of vertices which leads the algorithm to reject. Let $m = \sqrt{n}/2\phi^3$.

For a vertex $v \in M$, $P_v = \{u \in M : \mathcal{M}_I(u,v) > \mathcal{M}(u,v) + 2\sigma \text{ or } \mathcal{M}_I(u,v) < \mathcal{M}(u,v) - 2\sigma\}$, and X_v denote the indicator variable that $X_v = 1$ iff there exists a $u \in P_v$ in the first m samples. Then

$$\Pr[X_v] = 1 - \left(1 - \frac{|P_v|}{|S|}\right)^m.$$

Since $|P_v| \le \phi^4|S|/\sqrt{n}$, we have

$$(1-o(1))\frac{|P_v|m}{|S|} \le \Pr[X_v = 1] \le \frac{|P_v|m}{|S|}.$$

Let $X = \sum_{v\in S} X_v$ and $p = \sum_{v\in S} |P_v|$. $\frac{(1-o(1))pm}{|S|} \le \mathbb{E}[X] \le \frac{pm}{|S|}$.

Now we calculate the variance of X.

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sum_{u,v\in S} \mathbb{E}[X_u X_v] - \mathbb{E}[X]^2$$

Since X_us are 0-1 random variables,

$$\sum_{u\in S} \mathbb{E}[X_u^2] = \sum_{u\in S} \mathbb{E}[X_u] \le \sum_{u\in S} \frac{|P_u|m}{|S|} = \frac{pm}{|S|}$$

Let $P_{u,v} = P_u \cap P_v$, $P_{u,-v} = P_u \backslash P_{u,v}$. Note that

$$\begin{aligned}\mathbb{E}[X_u X_v] =& \Pr[X_u = 1, X_v = 1] \\ =& 1 - \Pr[X_u = 0, X_v = 1] - \Pr[X_u = 1, X_v = 0] - \Pr[X_u = 0, X_v = 0].\end{aligned}$$

Notice that

$$\Pr[X_u = 0, X_v = 0] = \left(1 - \frac{|P_{u,v}| + |P_{u,-v}| + |P_{v,-u}|}{|S|}\right)^m,$$

$$\Pr[X_u = 1, X_v = 0] = \left(1 - \frac{|P_{u,v}| + |P_{v,-u}|}{|S|}\right)^m - \left(1 - \frac{|P_{u,v}| + |P_{u,-v}| + |P_{v,-u}|}{|S|}\right)^m,$$

$$\Pr[X_u = 0, X_v = 1] = \left(1 - \frac{|P_{u,v}| + |P_{u,-v}|}{|S|}\right)^m - \left(1 - \frac{|P_{u,v}| + |P_{u,-v}| + |P_{v,-u}|}{|S|}\right)^m,$$

we have

$$\begin{aligned}\mathbb{E}[X_u X_v] =& 1 - \left(1 - \frac{|P_{u,v}| + |P_{v,-u}|}{|S|}\right)^m \\ & - \left(1 - \frac{|P_{u,v}| + |P_{u,-v}|}{|S|}\right)^m + \left(1 - \frac{|P_{u,v}| + |P_{u,-v}| + |P_{v,-u}|}{|S|}\right)^m.\end{aligned}$$

Since $\frac{(|P_{u,v}|+|P_{u,-v}|+|P_{v,-u}|)m}{|S|} = o(1)$,

$$\mathbb{E}[X_u X_v] \leq (1+o(1))\left(\frac{m|P_{u,v}|}{|S|} + \frac{m^2|P_{u,-v}||P_{v,-u}|)}{|S|^2}\right)$$

Thus,

$$\begin{aligned}\sum_{u\neq v\in M}\mathbb{E}(X_u X_v) \leq& (1+o(1))\left(\frac{m\sum_{u\neq v\in M}|P_{u,v}|}{|S|} + \frac{m^2\sum_{u\neq v\in M}|P_{u,-v}||P_{v,-u}|}{|S|^2}\right)\\ \leq& (1+o(1))\left(\frac{mp\phi^4|S|}{|S|\sqrt{n}} + \frac{m^2p^2}{|S|^2}\right)\end{aligned}$$

and finally

$$\begin{aligned}\mathrm{Var}[X] =& \mathbb{E}[X^2] - \mathbb{E}[X]^2\\ =& \sum_{u\neq v\in S}\mathbb{E}[X_u X_v] + \sum_{u\in S}\mathbb{E}[X_u^2] - \mathbb{E}[X^2]\\ \leq& (1+o(1))\left(\frac{mp\phi^4|S|}{|S|\sqrt{n}} + \frac{m^2p^2}{|S|^2}\right) + \frac{pm}{|S|} - \left(\frac{pm}{|S|}\right)^2\\ \leq& (1+o(1))\left(\frac{pm\phi^4}{\sqrt{n}} + \frac{pm}{|S|}\right) + o(\frac{m^2p^2}{|S|^2})\end{aligned}$$

Since $|P_u| \geq \frac{2\phi^3|S|^2}{|M|n}$ for every vertex in M, $p \geq 2\phi^3|S|^2/n$, which implies $\sqrt{\frac{pm\phi^4}{\sqrt{n}}} = o(\frac{pm}{|S|})$. Thus, $\sqrt{\mathrm{Var}(X)} = o(\mathbb{E}[X])$. By Chebyshev's inequality, we have

$$\Pr[X \geq |S|/2\sqrt{n}] \leq \Pr[X \geq \mathbb{E}[X]/2] = 1 - o(1).$$

On the other hand, if $X \geq |S|/2\sqrt{n}$, the probability that there is no sampled vertex v in the second $\sqrt{n}/2\phi^2$ samples satisfying (6.2) for some vertex $u \in M$ in the first $\sqrt{n}/2\phi^2$ samples is at least

$$\left(1 - \frac{1}{2\sqrt{n}}\right)^{\sqrt{n}/2\phi^2} = \exp(-\Omega(1/\phi^2)).$$

Overall, a random sequence of $\sqrt{n}/\phi^3$ samples can find a pair of vertices satisfying (6.2) with at least constant probability. Thus, a random sequence of $\sqrt{n}\log(1/\delta)/\phi^{16}$ samples can find a pair of vertices satisfying (6.2) with probability at least $1-\delta$.

Using above techniques, we can show that if the third or fourth condition of Definition 6.4.10, then Subroutine `Testing-Distance-Preserved-Subset` rejects with probability at least $1-\delta$. □

6.4.2 An efficient algorithm for testing collision problem

We prove Theorem 6.4.5 in this section. We reduce the Testing-Collision problem to the following problem.

Testing-Subset-Collision Problem

Input: $0 \leq i, j, k \leq 6 \log n/\varepsilon_1$ such that $\mathsf{weight}_w : \cup T_{i,j,k} \to \mathbb{R}^{\geq 0}$ satisfying $\mathsf{weight}_w(T_{i,j,k}) \geq \varepsilon_2^2 n$, parameter δ.

Output:

1. Accept with probability $1 - \delta$ if all of the following conditions hold
   (a) Every pair of vertices $v, w \in T_{i,j,k}$ satisfies $\mathcal{M}_I(v, w) - 2\sigma \leq \mathcal{M}(v, w) \leq \mathcal{M}_I(v, w) + 2\sigma$.
   (b) Every pair of vertices y, z in $H_{i,j,k}$ or $C_{i,j,k}$ satisfies $\mathcal{M}_J(y, z) - 2\sigma \leq \mathcal{M}(y, z) \leq \mathcal{M}_J(y, z) + 2\sigma$.
   (c) For every $x \in T_{i,j,k}, y \in C_{i,j,k}$ satisfying $\mathcal{M}(x, y) \leq r$, y is a useful collision of x.
   (d) Every vertex $x \in T_{i,j,k}$, x has at least $(1 - \varepsilon_1)B_I(x, r)$ and at most $(1 + \varepsilon_1)B_I(x, r)$ useful collisions.
2. Reject with probability $1 - \delta$ if at least one of following conditions hold
   (a) At least $|S_{i,j,k}|/\sqrt{n}$ vertices $x \in S_{i,j,k}$ have distance distorted to $Assignto(x)$.
   (b) At least $\phi^4 |H_{i,j,k}|/\sqrt{n}$ vertices are at least ϕ^2-distorted in $H_{i,j,k}$ by f_J.
   (c) A total weight of at least $10\varepsilon_1 \mathsf{weight}_w(T_{i,j,k})$ vertices in $T_{i,j,k}$ do not have semi-useful collision.

We start with some useful subroutines.

Subroutine `Testing-Good-Collision`**:**

Input: Two vertices $v \in T_{i,j,k}$ and $y \in C_{i,j,k}$ satisfying $\mathcal{M}(v, y) \leq r$, a parameter δ.

Output: Accept or reject.

1. If there is no $z \in V_J$ satisfying $\mathcal{M}_J(y,z) \leq r+2\sigma$ and $\rho_4(x,z) \leq 2\sigma$, then rejects.
2. Set $c_1 = 0$ and $c_2 = 0$.
3. Randomly sample $m = \lceil 2\log(1/\delta)/\varepsilon_1^2\phi \rceil$ vertices in $B_J(y, 2r+6\sigma)$. For every sampled vertex z satisfying $\mathcal{M}(v,z) \leq r$, if there is a vertex $u \in V_J$ satisfying $\mathcal{M}_J(z,u) \leq r+2\sigma$ and $\rho_4(x,u) \leq 2\sigma$, then increase c_1 by 1, otherwise reject.
4. Randomly sample m vertices z in $B_J(y, 2r+6\delta, 29r)$, if $\mathcal{M}(v,z) \leq r$ then rejects.
5. If $c_1 > \frac{(1+1.5\varepsilon_1)|B_I(x,r)|}{|B_J(y,2r+6\sigma)|}m$ or $c_1 < \frac{(1-1.5\varepsilon_1)|B_I(x,r)|}{|B_J(y,2r+6\sigma)|}m$, then rejects, otherwise accepts.

Remark 6.4.12. *Given two vertices $v \in T_{i,j,k}$ and $y \in C_{i,j,k}$ with $\mathcal{M}(v,y) \leq r$. By the definition of collision, it is possible to determine whether y is a collision of v without checking f_J of other vertices in V_J.*

Lemma 6.4.13. *If y is not a collision with v, then with probability at least 1, Subroutine* `Testing-Good-Collision` *rejects.*

If y is a good collision with v, then with probability at least $1-\delta$, Subroutine `Testing-Good-Collision` *accepts. If y is a bad collision with v, then with probability at least $1-\delta$, Subroutine* `Testing-Good-Collision` *rejects.*

Proof. If y is not a collision to v, then the subroutine rejects with probability 1 at Step 1.

Let X_i denote the indicator variable whether i-th sample in Step 3 is a collision to x. Then $\Pr[X_i] = \frac{|N(v,y,2r+6\delta)|}{|B_J(y,2r+6\sigma)|}$, and $\mathbb{E}[X = \sum X_i] = \frac{|N(v,y,2r+6\delta)|m}{|B_J(y,2r+6\sigma)|}$. Since y is a collision to v, there exists a vertex z satisfying $\mathcal{M}_J(y,z) \leq r+2\sigma$ and $\rho_4(v,z) \leq 2\sigma$. Thus $|B_I(v,r)| \geq 2\phi|B_I(v,30r+6\sigma)| \geq 2\phi|B_J(z,30r+4\sigma)|$. On the other hand, $B_J(y,29r) \subseteq B_J(z,30r+4\sigma)$, so we have $|B_I(v,r)| \geq 2\phi|B_J(y,29r)|$ and thus $|B_I(v,r)| \geq 2\phi|B_J(y,2r+6\sigma)|$.

If y is a good collision to v, then

$$(1-\varepsilon_1)\phi m \leq \frac{(1-\varepsilon_1)|B_I(v,r)|m}{|B_J(y,2r+6\sigma)|} \leq \mathbb{E}[X] \leq \frac{(1+\varepsilon_1)|B_I(v,r)|m}{|B_J(y,2r+6\sigma)|}.$$

By Chernoff bound, with probability at least $1-\delta$, the subroutine accepts.

If y is a bad collision, then one of following conditions hold:

1. $\mathbb{E}[X] < \frac{(1-2\varepsilon_1)|B_I(v,r)|m}{|B_J(y,2r+6\sigma)|}$ or $\mathbb{E}[X] > \frac{(1+2\varepsilon_1)|B_I(v,r)|m}{|B_J(y,2r+6\sigma)|}$
2. The probability that a random vertex z in Step 3 is not a collision of x but satisfying $\mathcal{M}(x,z) \leq r$ is at least $\frac{\varepsilon_1|B_I(x,r)|}{|B_J(y,2r+6\delta)|}$.
3. The probability that a random vertex z in Step 4 is not a collision of x but satisfying $\mathcal{M}(x,z) \leq r$ is at least $\frac{\varepsilon_1|B_I(x,r)|}{|B_J(y,2r+6\delta,29r)|}$.

Hence, with probability at least $1-\delta$, the subroutine rejects. □

Subroutine `Testing-Useful-Collision`**:**

Input: Two vertices $v \in T_{i,j,k}$ and $y \in C_{i,j,k}$ satisfying y is a collision to v, a parameter δ

Output: Accept or reject.

1. Run Subroutine `Testing-Good-Collision` with vertex v and y and parameter $\delta/2$. If `Testing-Good-Collision` rejects, then rejects.
2. Randomly sample $m = \lfloor 2\log(1/\delta)/\varepsilon_1\phi \rfloor$ vertices z in $B_J(y, 2r+6\sigma)$: If z is a collision of v, then run Subroutine `Testing-Good-Collision` with vertex v and z and parameter $\delta/4m$. If any execution of Subroutine `Testing-Good-Collision` rejects, then rejects.
3. Accepts.

Lemma 6.4.14. *If y is a useful collision with v, then Subroutine* `Testing-Useful-Collision` *accepts with probability at least $1-\delta$. If y is not a collision with v or is a useless collision with v, then Subroutine* `Testing-Useful-Collision` *rejects with probability at least $1-\delta$.*

Proof. By Lemma 6.4.13, if y is a bad collision to v, then Step 1 of Subroutine `Testing-Useful-Collision` rejects with probability at least $1-\delta/2$.

Now we assume y is a good or intermediate collision to v. The number of collisions to v in $B_J(y, 2r + 6\sigma)$ is between $(1 - 2\varepsilon_1)|B_I(v, r)|$ and $(1 + 2\varepsilon_1)|B_I(v, r)|$. If y is a useless collision to v, then at least $\varepsilon_1|B_I(v, r)|$ of them are bad. With probability

$$1 - (1 - \frac{\varepsilon_1|B_I(v,r)|}{|B_J(y,2r+6\sigma)|})^m \geq 1 - (1 - \varepsilon_1\phi)^m = 1 - \delta/4,$$

at least one sampled vertex is a bad collision of v. By Lemma 6.4.13, Subroutine `Testing-Single-Collision` rejects with probability at least $1-\delta/4m$. By union bound, Subroutine `Testing-Useful-Collision` rejects with probability at least $1 - \delta$.

If y is a useful collision to v, then by Lemma 6.4.13 and union bound, Subroutine passes Step 2 with probability at least $1 - \delta$. □

Subroutine `Testing-Random-Subset-Collision`**:**

Input: $S_{i,j,k}, T_{i,j,k}$ for $0 \leq i, j, k \leq 6 \log n/\varepsilon_1$. A set of vertices $Q \subseteq T_{i,j,k}$ such that $|Q| \geq \phi^8|T_{i,j,k}|/\sqrt{n}$ and every pair of vertices in Q has distance at least $4r + 4\sigma$ in $\mathcal{M}$, a set of vertices $C_{i,j,k}$, and a parameter δ.

Output: Accept or reject. A pair of vertices (v, y) with $v \in S_{i,j,k}$, $y \in C_{i,j,k}$.

1. Randomly sample a set K of $\lceil \frac{96\sqrt{n}\log(1/\delta)}{\phi^{20}\varepsilon_2^2\varepsilon_1^2} \rceil$ vertices in $C_{i,j,k}$. Randomly select a pair of vertices $(x \in Q, z \in K)$ satisfying $\mathcal{M}(x, z) \leq r$, and randomly sample a vertex $v \in Assign(x)$. Set $p = 0$.

2. For every vertex $y \in K$, if there exists a vertex $v \in Q$ such that $\mathcal{M}(v, y) \leq r$, then run Subroutine `Testing-Useful-Collision` with vertices y, v and parameters δ/n^2. If the subroutine accepts, then increase p by 1.

3. Let $q = \frac{|Q||K|\gamma_{i,j,k}}{|C_{i,j,k}|}$. If $p < q(1-1.5\varepsilon_1)$, $p > q(1+\varepsilon_1)(1+1.5\varepsilon_1)$ or any execution of Subroutine `Testing-Useful-Collision` in Step 1 or Step 2 rejects, then reject, otherwise accept. Return (v, z).

Given a subset Q of $T_{i,j,k}$, define bipartite graph $A_u(Q) = (Q, C_{i,j,k})$ such that there is an edge between $x \in Q$ and $y \in C_{i,j,k}$ if and only if y is a useful collision to x, bipartite

graph $A_s(Q) = (Q, C_{i,j,k})$ such that there is an edge between $x \in Q$ and $y \in C_{i,j,k}$ if and only if y is a semi-useful collision to x, and bipartite graph $A_n(Q) = (Q, C_{i,j,k})$ such that there is an edge between $x \in Q$ and $y \in C_{i,j,k}$ if $\mathcal{M}(x, y) \leq r$ and y is a not a semi-useful collision to x. We use $|A_u(Q)|$ to denote the number of edges in bipartite graph $A_u(Q)$ (and same notation for all the following bipartite graphs).

Lemma 6.4.15. *If for every $x \in Q, y \in C_{i,j,k}$ with $\mathcal{M}(x, y) \leq r$ y is a useful collision of x, and $(1-\varepsilon_1)|Q|\gamma_{i,j,k} \leq |A_u(Q)| \leq (1+\varepsilon_1)^2|Q|\gamma_{i,j,k}$, then the subroutine* `Testing-Random-Subset-Collision` *accepts with probability at least $1-\delta$.*

If $|A_n(Q)| > \varepsilon_1^2\phi^{10}|Q|\gamma_{i,j,k}$, $|A_s(Q)| < (1-2\varepsilon_1)|Q|\gamma_{i,j,k}$ or $|A_u(Q)| > (1+\varepsilon_1)(1+2\varepsilon_1)|Q|\gamma_{i,j,k}$, then the subroutine `Testing-Random-Subset-Collision` *rejects with probability at least $1-\delta$.*

Proof. Since for every $v, w \in Q$ satisfying $\mathcal{M}(v, w) \geq 4r + 4\sigma$, every vertex in $C_{i,j,k}$ can be a collision of at most one vertex in Q. Let X_i, Y_i and Z_i be the indicator variable whether i-th sampled vertex y in step 2 of Subroutine `Testing-Random-Subset-Collision` is a useful collision, semi-useful or not semi-useful collision but with distance at most r to some vertex in Q respectively. Hence $\Pr[X_i] = \frac{|A_u(Q)|}{|C_{i,j,k}|}$, $\Pr[Y_i] = \frac{|A_s(Q)|}{|C_{i,j,k}|}$ and $\Pr[Z_i] = \frac{|A_n(Q)|}{|C_{i,j,k}|}$. Thus, $\mathbb{E}[X = \sum X_i] = \frac{|K||A_u(Q)|}{|C_{i,j,k}|}$, $\mathbb{E}[Y = \sum Y_i] = \frac{|K||A_s(Q)|}{|C_{i,j,k}|}$ and $\mathbb{E}[Z = \sum Z_i] = \frac{|K||A_n(Q)|}{|C_{i,j,k}|}$.

By Corollary 6.3.5, we have

$$\frac{|K||Q|\gamma_{i,j,k}}{|C_{i,j,k}|} \geq \frac{|K|\phi^8|T_{i,j,k}|\gamma_{i,j,k}\varepsilon_2^2\phi}{4\sqrt{n}|T_{i,j,k}|\gamma_{i,j,k}} \geq \frac{24\log(1/\delta)}{\phi^{11}\varepsilon_1^2}.$$

If $(1-\varepsilon_1)|Q|\gamma_{i,j,k} \leq |A_u(Q)| \leq (1+\varepsilon_1)^2|Q|\gamma_{i,j,k}$, then

$$(1-\varepsilon_1)|K||Q|\gamma_{i,j,k}/|C_{i,j,k}| \leq \mathbb{E}[X] \leq (1+\varepsilon_1)^2|K||Q|\gamma_{i,j,k}/|C_{i,j,k}|.$$

By Chernoff bound, we have

$$\Pr[(1-1.5\varepsilon_1)q \leq X \leq (1+\varepsilon_1)(1+1.5\varepsilon_1)q] \geq 1-\delta,$$

and thus if every collision for vertices in Q is useful and $(1-\varepsilon_1)|Q|\gamma_{i,j,k} \leq |A_u(Q)| \leq (1+\varepsilon_1)^2|Q|\gamma_{i,j,k}$, then the subroutine `Testing-Random-Subset-Collision` accepts with probability at least $1-\delta$.

Similarly, If $|A_n(Q)| > \varepsilon_1^2\phi^{10}|Q|\gamma_{i,j,k}$, $|A_s(Q)| < (1-2\varepsilon_1)|Q|\gamma_{i,j,k}$ or $|A_u(Q)| > (1+\varepsilon_1)(1+2\varepsilon_1)|Q|\gamma_{i,j,k}$, then at least one of following three conditions hold:

$$\mathbb{E}[Y] < \frac{(1-2\varepsilon_1)|K||Q|\gamma_{i,j,k}}{|C_{i,j,k}|}, \mathbb{E}[X] > \frac{(1+\varepsilon_1)(1+2\varepsilon_1)|K||Q|\gamma_{i,j,k}}{|C_{i,j,k}|},$$
$$\mathbb{E}[Z] > \frac{\varepsilon_1^2\phi^{10}|K||Q|\gamma_{i,j,k}}{|C_{i,j,k}|}.$$

Again, by Chernoff bound, we obtain the lemma. □

Subroutine `Testing-Subset-Collision`**:**

Input: $S_{i,j,k}, T_{i,j,k}$ and $C_{i,j,k}$ for $0 \le i,j,k \le 6\log n/\varepsilon_1$ with $\mathsf{weight}_w(T_{i,j,k}) \ge \varepsilon_2^2 n$.

Output: Accept or reject. A pair of vertices (v,y) with $v \in S_{i,j,k}$, $y \in C_{i,j,k}$.

1. Randomly sample a set W of $2\sqrt{n}\log(2/\delta)$ vertices in $S_{i,j,k}$. If there is $v \in W$ such that $\mathcal{M}_I(v, Assignto(v)) > \mathcal{M}(v, Assignto(v)) + 2\sigma$ or $\mathcal{M}_I(v, Assignto(v)) < \mathcal{M}(v, Assignto(v)) - 2\sigma$, then reject, and return an empty pair.

2. Run Subroutine `Testing-Distance-Preserved-Subset` on $T_{i,j,k}$, $S_{i,j,k}$ and $C_{i,j,k}$ with uniform weight function and parameter δ/n. If any execution of the subroutine rejects, then reject, and return an empty pair.

3. If $|T_{i,j,k}| \le \sqrt{n}/\phi^{14}$, then let $Q = T_{i,j,k}$

   (a) Reject and return an empty pair if there exists $v, w \in Q$ such that $\mathcal{M}_I(v,w) > \mathcal{M}(v,w) + 2\sigma$ or $\mathcal{M}_I(v,w) < \mathcal{M}(v,w) - 2\sigma$.

   (b) Randomly sample a set of $48\sqrt{n}\log(2n/\delta)/\varepsilon_2^2\phi^{15}$ vertices K in $C_{i,j,k}$. For every $x \in Q$ and $v \in K$ satisfying $\mathcal{M}(x,v) \le r$, run subroutine `Testing-Useful-Collision` with vertex x, v and parameter $\delta/2n^2$.

   (c) Reject if any execution of subroutine `Testing-Useful-Collision` rejects or there is a vertex $v \in Q$ such that less than $\frac{(1-1.5\varepsilon_1)|B_I(v,r)||Q|}{|C_{i,j,k}|}$ or more than $\frac{(1+1.5\varepsilon_1)|B_I(v,r)||Q|}{|C_{i,j,k}|}$ vertices in K have distance less than r in $\mathcal{M}$, otherwise, accept.

(d) Randomly select a vertex $t \in Q$. Randomly choose a $u \in Assign(t)$ and $z \in K$ satisfying $\mathcal{M}(t,z) \leq r$. Return the pair (u,z).

4. If $|T_{i,j,k}| > \sqrt{n}/\phi^{14}$, then let $P = \emptyset$ initially and repeat following process $\log(4/\delta)/\phi^8$ times:

   (a) Randomly sample a set of $\lceil \phi^8 |T_{i,j,k}|/\sqrt{n} \rceil$ vertices in $T_{i,j,k}$ (with replacement). Let Q denote this set of vertices.

   (b) Reject and return an empty pair if there exists $v, w \in Q$ such that $\mathcal{M}_I(v,w) > \mathcal{M}(v,w) + 2\sigma$ or $\mathcal{M}_I(v,w) < \mathcal{M}(v,w) - 2\sigma$.

   (c) Run Subroutine `Testing-Random-Subset-Collision` with Q and parameter δ/n. If the subroutine rejects, then reject, otherwise, put the pair returned into P.

5. Randomly select a pair $(v,y) \in P$. Randomly select a vertex $x \in Assign(v)$. Accept and return (x,y).

Lemma 6.4.16. *If $|T_{i,j,k}| > \sqrt{n}/\phi^{14}$, then every execution of step 4(a) of Subroutine* `Testing-Subset-Collision` *obtain distinct vertex samples with probability at least $1 - O(\phi^{16})$, and at least half of the executions of step 4(a) obtain distinct vertex samples with probability at least $1 - \delta/4$, .*

Proof. Let X_i be the indicator variable that whether the i-th vertex sample is distinct to all the previous vertex samples. We have $\Pr[X_i] \geq 1 - \frac{|Q|}{|T_{i,j,k}|}$. Thus, the probability that all the vertex samples are distinct is at least

$$\left(1 - \frac{|Q|}{|T_{i,j,k}|}\right)^{|Q|} \geq \left(1 - O\left(\frac{\phi^8}{\sqrt{n}}\right)\right)^{\phi^8 |T_{i,j,k}|/\sqrt{n}} \geq 1 - O\left(\frac{\phi^{16}|T_{i,j,k}|}{n}\right) \geq 1 - O(\phi^{16}). \quad (6.3)$$

By Chernoff bound, we obtain the lemma. □

Lemma 6.4.17. *If $|T_{i,j,k}| > \sqrt{n}/\phi^{14}$ and $|A_s(T_{i,j,k})| < (1-4\varepsilon_1)|T_{i,j,k}|\gamma_{i,j,k}$, then Subroutine* `Testing-Subset-Collision` *rejects with probability $1 - \delta$.*

Proof. Let X_i denote the number of semi-useful collisions in $C_{i,j,k}$ to the i-th vertex sample in Step 4(a) of Subroutine `Testing-Subset-Collision`. We have $\mathbb{E}[X_i] = \frac{|A_s(T_{i,j,k})|}{|T_{i,j,k}|}$, and thus $\mathbb{E}[X = \sum X_i] = \frac{|A_s(T_{i,j,k})||Q|}{|T_{i,j,k}|}$. On the other hand, let Y be the event that all the vertex samples in Step 4(a) are distinct, we have

$$\mathbb{E}[X] \geq \Pr[Y]\,\mathbb{E}[X|Y] = \Pr[Y]\,\mathbb{E}[|A_s(Q)||Y] \geq (1 - O(\phi^{16}))\,\mathbb{E}[|A_s(Q)||Y].$$

Thus,

$$\mathbb{E}[|A_s(Q)||Y] \leq \frac{|A_s(T_{i,j,k})||Q|}{|T_{i,j,k}|(1 - O(\phi^{16}))} \leq \frac{(1 - 4\varepsilon_1)\gamma_{i,j,k}|Q|}{1 - O(\phi^{16})} \leq (1 - 3\varepsilon_1)\gamma_{i,j,k}|Q|.$$

Now we show that if $\mathbb{E}[|A_s(Q)||Y] < (1-3\varepsilon_1)|Q|\gamma_{i,j,k}$, then Subroutine `Testing-Subset-Collision` rejects with probability at least $1 - \exp(-\Omega(1/\phi^8))$. If $\mathbb{E}[|A_s(Q)||Y] < (1 - 3\varepsilon_1)|Q|\gamma_{i,j,k}$, then by Markov inequality,

$$\Pr[|A_s(Q)| \geq (1 - 2\varepsilon_1)|Q|\gamma_{i,j,k}|Y] \leq \frac{\mathbb{E}[|A_s(Q)||Y]}{(1 - 2\varepsilon_1)|Q|\gamma_{i,j,k}} < 1 - \frac{\varepsilon_1}{1 - 2\varepsilon_1}.$$

By Lemma 6.4.16, with probability at least $1 - \delta/4$, there is one execution of Step 4(a) obtaining distinct vertex samples, and $|A_s(Q)| < (1 - 2\varepsilon_1)|Q|\gamma_{i,j,k}$. By Lemma 6.4.15, Algorithm `Testing-Subset-Collision` rejects with probability at least $1 - \delta/2$. □

Let W_x be a set of collisions of vertex x satisfying

1. Every vertex y in W_x is a good/intermediate collision with x.
2. For every pair of vertices $y, z \in W_x$, we have $\mathcal{M}_J(y, z) \geq 29r$.
3. The size of W_x is maximized.

Let $L_{i,j,k}$ be the set of vertices $x \in T_{i,j,k}$ satisfying $|W_x| \geq 2$, and

$$R_{i,j,k} = \{y \in C_{i,j,k} : \exists x \in L_{i,j,k} \text{ s.t. } y \text{ is a good or intermediate collision of } x\}.$$

We define following bipartite graph $Y = (L_{i,j,k}, C_{i,j,k})$, in which there is an edge between $x \in L_{i,j,k}$ and $y \in C_{i,j,k}$ if and only if y is a good/intermediate collision to x.

We prove following lemma.

Lemma 6.4.18. *If $|T_{i,j,k}| > \sqrt{n}/\phi^{14}$ and $|Y| \geq \phi|T_{i,j,k}|\gamma_{i,j,k}$, then either $T_{i,j,k}$ or $C_{i,j,k}$ have a distorted subset.*

We need some more definitions. For $x \in L_{i,j,k}$, $y \in R_{i,j,k}$, let $P_{x,y} = \{z \in B_J(y, 2r+6\delta) : z \text{ is a collision of } x\}$ if y is a good/intermediate collision of x, otherwise $P_{x,y}$ is an empty set. Let $X_{i,j,k} = \cup_{x \in L_{i,j,k}, y \in R_{i,j,k}} P_{x,y}$. Let $\widehat{Y}$ be the bipartite graph on $(L_{i,j,k}, X_{i,j,k})$ such that $x \in L_{i,j,k}$ is adjacent to $y \in \cup_{x \in L_{i,j,k}}$ iff there is a $z \in R_{i,j,k}$ such that $y \in P_{x,z}$.

Let Z be a subgraph of $\widehat{Y}$. For any $x \in L_{i,j,k}$, let

$$D_Z(x) = \{v \in L_{i,j,k} : \exists y \in X_{i,j,k} \text{ s.t. } y \text{ is adjacent to both } x \text{ and } v \text{ in } Z\}.$$

For any $y \in X_{i,j,k}$, let

$$\begin{aligned} D_Z(y) = \{z \in X_{i,j,k} :& \exists x \in L_{i,j,k}, u \in X_{i,j,k} \text{ s.t. } y \in P_{x,u}, \\ & z \text{ is adjacent to } x \text{ in } Z \text{ and } \mathcal{M}_J(u,z) > 29r\}. \end{aligned}$$

Fact 6.4.19. *Let Z be a subgraph of $\widehat{Y}$. For any $x \in L_{i,j,k}$ and $v, t \in \{x\} \cup D_Z(x)$, the distance between v and t is distorted by f_I. For any $y \in X_{i,j,k}$ and $z \in D_Z(y)$, the distance between y and z is distorted by f_J.*

Lemma 6.4.20. *Let Z be a subgraph of $\widehat{Y}$. If $x \in L_{i,j,k}$ and $y \in X_{i,j,k}$ with $(x,y) \in Z$, then $|D_Z(y)| \geq \deg_Z(x) - (1+3\varepsilon_1)(1+\varepsilon_1)\gamma_{i,j,k}$ and $|D_Z(x)| \geq \deg_Z(y) - 1$ hold, where $\deg_Z(x)$ denote the degree of vertex x in graph Z.*

Proof. By the definition of $\widehat{Y}$, there is a $z \in R_{i,j,k}$ such that z is a good/intermediate collision of x, and $y \in B_J(z, 2r+6\delta)$. By Fact 6.4.2, there are at most $(1+3\varepsilon_1)(1+\varepsilon_1)\gamma_{i,j,k}$ vertices $u \in B_J(y, 29r)$ satisfying that u is a collision of x. So, at least $\deg_Z(x) - (1+3\varepsilon_1)(1+\varepsilon_1)\gamma_{i,j,k}$ vertices in Z have distance distorted with y.

On the other hand, if both x and x' are adjacent to y in Z, then the distance between x and x' is distorted. So, at least $\deg_Z(y) - 1$ vertices in Z have distance distorted with x. □

Proof of Lemma 6.4.18. Since every vertex $x \in L_{i,j,k}$ satisfies $|W_x| \geq 2$, we can find a ℓ with $\ell \geq 2(1-3\varepsilon_1)\gamma_{i,j,k}$ such that

$$\sum_{v \in L_{i,j,k} : \ell \leq \deg_{\widehat{Y}}(v) < 2\ell} \deg_{\widehat{Y}}(v) \geq \frac{|\hat{Y}|}{\log_2 n} \geq \frac{|Y|}{\log_2 n} \geq \frac{\phi |T_{i,j,k}| \gamma_{i,j,k}}{\log_2 n},$$

where $\deg_{\widehat{Y}}(v)$ is the degree of vertex v in bipartite graph $\widehat{Y}$. Let $L'_{i,j,k} = \{v \in L_{i,j,k} : \ell \leq \deg_{\widehat{Y}}(v) < 2\ell\}$, $X'_{i,j,k} = \{z \in X_{i,j,k} : |N_{\widehat{Y}}(z) \cap L'_{i,j,k}| \geq 1\}$, where $N_{\widehat{Y}}(z)$ denote the set of neighbors of z in graph $\widehat{Y}$. Let $\widehat{Y}'$ be the induced subgraph of $(L'_{i,j,k}, X'_{i,j,k})$ in $\widehat{Y}$. Furthermore, let $AveL = \frac{|\widehat{Y}'|}{|L'_{i,j,k}|}$ and $AveX = \frac{|\widehat{Y}'|}{|X'_{i,j,k}|}$.

If $|X'_{i,j,k}| \geq 16|L'_{i,j,k}|/\log_2^{10} n$, then by Lemma 6.4.20, for every vertex $y \in X'_{i,j,k}$, we have

$$|D_{\widehat{Y}'}(y)| \geq \max\{1, \ell - (1+3\varepsilon_1)(1+\varepsilon_1)\gamma_{i,j,k}\} \geq \frac{(1-6\varepsilon_1)\ell}{2}.$$

On the other hand, since $|\widehat{Y}'| < 2\ell|L'_{i,j,k}|$,

$$|D_{\widehat{Y}'}(y)| \geq \frac{(1-6\varepsilon_1)\ell}{2} > \frac{|\widehat{Y}'|}{8|L'_{i,j,k}|} \geq \frac{\phi|T_{i,j,k}|\gamma_{i,j,k}}{8\log_2 n|L'_{i,j,k}|} \geq \frac{2\phi|T_{i,j,k}|\gamma_{i,j,k}}{\log_2^{11} n|X'_{i,j,k}|} \geq \frac{2\phi^3|C_{i,j,k}|}{|X'_{i,j,k}|}, \quad (6.4)$$

where the last inequality uses Corollary 6.3.5.

If there exists a vertex $y \in X'_{i,j,k}$ with $|D_{\widehat{Y}'}(y)| \geq \phi^4|C_{i,j,k}|/\sqrt{n}$, then by Corollary 6.3.5, for any vertex $z \in \{y\} \cup D_{\widehat{Y}'}(y)$, at least $|D_{\widehat{Y}'}(y)| - \frac{(1+\varepsilon_1)\gamma_{i,j,k}}{2\phi} = (1-o(1))|D_{\widehat{Y}'}(y)|$ vertices in $\{y\} \cup D_{\widehat{Y}'}(y)$ has distance distorted to z. Hence,$\{y\} \cup D_{\widehat{Y}'}(y)$ forms a distorted set in $C_{i,j,k}$. Otherwise, for an arbitrary $y \in X'_{i,j,k}$, we have

$$|X'_{i,j,k}| \geq \frac{2\phi^3|C_{i,j,k}|}{|D_{\widehat{Y}'}(y)|} > \frac{2\phi^3|C_{i,j,k}|\sqrt{n}}{\phi^4|C_{i,j,k}|} = \frac{2\sqrt{n}}{\phi},$$

and by Inequality (6.4), $X'_{i,j,k}$ is a distorted set in $C_{i,j,k}$.

Now we consider the case of $|X'_{i,j,k}| < 16|L'_{i,j,k}|/\log_2^{10} n$. Let $L''_{i,j,k} = L'_{i,j,k}$ and $X''_{i,j,k} = X'_{i,j,k}$ initially, and let $\widehat{Y}'' = (L''_{i,j,k}, X''_{i,j,k})$. We keep following process on $\widehat{Y}''$ until there is no more action available:

1. Remove vertex $x \in L''_{i,j,k}$ from $\widehat{Y}''$ if the degree of x is smaller than $AveL/4$
2. Remove vertex $y \in X''_{i,j,k}$ from $\widehat{Y}''$ if the degree of y is smaller than $AveX/4$.

Then every vertex in $L''_{i,j,k}$ has degree at least $AveL/4$ in $\widehat{Y}''$, and every vertex in $X''_{i,j,k}$ has degree at least $AveX/4$ in $\widehat{Y}''$. On the other hand, The total number of edges remaining in $\widehat{Y}''$ is at least

$$|\widehat{Y}'| - \frac{|L'_{i,j,k}|AveL}{4} - \frac{|X'_{i,j,k}|AveX}{4} = |\widehat{Y}'|/2.$$

So, every vertex $x \in L''_{i,j,k}$ has $|D_{\widehat{Y}''}(x)| \geq \frac{AveX}{4} - 1$. Notice that

$$\begin{aligned}
|D_{\widehat{Y}''}(x)| \geq & \frac{AveX}{4} - 1 \\
= & \frac{|\widehat{Y}'|}{4|X'_{i,j,k}|} - 1 \\
\geq & \frac{|\widehat{Y}'|}{64|L'_{i,j,k}|/\log_2^{10} n} - 1 \\
\geq & \frac{\log_2^{10} n\ell}{64} - 1 \\
= & \frac{(1-o(1)\ell \log_2^{10} n}{64}.
\end{aligned}$$

If there exists a vertex $x \in L''_{i,j,k}$ such that $|D_{\widehat{Y}''}(x)| > \phi^4 |T_{i,j,k}|/\sqrt{n}$, then by Fact 6.4.19, $D_{\widehat{Y}''}(x)$ is a distorted set in $T_{i,j,k}$. Otherwise, since $|L''_{i,j,k}| \geq \frac{|\widehat{Y}'|}{4\ell} \geq \frac{\phi|T_{i,j,k}|\gamma_{i,j,k}}{4\log_2 n\ell}$, we have

$$|D_{\widehat{Y}''}(x)||L''_{i,j,k}| \geq \frac{(1-o(1))\ell \log_2^{10} n}{64}|L''_{i,j,k}| \geq \frac{\phi \log_2^9 n|T_{i,j,k}|\gamma_{i,j,k}}{512} > \frac{2\phi^3|T_{i,j,k}|^2}{n}.$$

Hence, $L''_{i,j,k}$ is a distorted set in $T_{i,j,k}$. □

Lemma 6.4.21. *If $|T_{i,j,k}| \leq \sqrt{n}/\phi^{14}$ and every pair of vertices in $T_{i,j,k}$ is not distorted, then*

1. *If Y is not an empty graph, then Subroutine* `Testing-Subset-Collision` *rejects with probability at least $1-\delta$.*

2. *If $A_n(T_{i,j,k}) > \phi^{15}\varepsilon_2^2\sqrt{n}$, then Subroutine* `Testing-Subset-Collision` *rejects with probability at least $1-\delta$.*

3. *If all the collisions are useful, and for every vertex $x \in T_{i,j,k}$, $(1-\varepsilon_1)|B_I(x,r)| \leq |A_u(x)| \leq (1+\varepsilon_1)|B_I(x,r)|$, then Subroutine* `Testing-Subset-Collision` *accepts with probability at least $1-\delta$.*

4. *If there is a vertex $x \in T_{i,j,k}$ satisfying $|A_s(x)| < (1-2\varepsilon_1)|B_I(x,r)|$ or $|A_s(x) \cup A_n(x)| > (1+2\varepsilon_1)|B_I(x,r)|$, then Subroutine* `Testing-Subset-Collision` *rejects with probability at least $1-\delta$.*

Proof. Assume there is a $x \in T_{i,j,k}$ with $|W_x| \geq 2$. Then there are two sets of vertices M_1 and M_2 with $|M_1|, |M_2| \geq (1-2\varepsilon_1)\gamma_{i,j,k}$ such that the distance between every vertex of

M_1 and every vertex of M_2 is distorted. Hence $M_1 \cup M_2$ is a distorted set of $C_{i,j,k}$. By Lemma 6.4.11, the subroutine rejects with probability at least $1 - \delta$.

By Lemma 6.4.14, the last three conditions hold. □

Lemma 6.4.22. *Subroutine* `Testing-Subset-Collision` *solve the Testing-Subset-Collision problem with probability at least* $1 - \delta$.

Proof. We first show that for a positive instance of the Testing-Subset-Collision problem, the subroutine accept with probability $1 - \delta$. Since any pair of vertices in $S_{i,j,k}, H_{i,j,k}$ or $C_{i,j,k}$ has distance not distorted, the first two steps pass with probability 1. For the case of $|T_{i,j,k}| \leq \sqrt{n}/\phi^{14}$, by Lemma 6.4.21, the subroutine accepts with probability at least $1 - \delta$. For the case of $|T_{i,j,k}| > \sqrt{n}/\phi^{14}$, every set Q sampled in Step 4(a) satisfying $(1 - \varepsilon_1)|Q|\gamma_{i,j,k} \leq |A_u(Q)| \leq (1 + \varepsilon_1)^2|Q|\gamma_{i,j,k}$. By Lemma 6.4.15, the subroutine accepts with probability at least $1 - \delta$.

Now we consider a negative instance of the Testing-Subset-Collision problem. If there are at least $|S_{i,j,k}|/\sqrt{n}$ vertices x in $S_{i,j,k}$ with distance distorted to $Assignto(x)$, then with probability at least $1 - \delta/2$, the subroutine rejects at Step 1. By Lemma 6.4.9, if at least $\phi^4|H_{i,j,k}|/\sqrt{n}$ vertices are at least ϕ^2-distorted in $H_{i,j,k}$ by f_J, then with probability at least $1 - \delta/2$, the subroutine rejects at Step 2.

Consider the case that a total weight of at least $10\varepsilon_1 \mathsf{weight}_w(T_{i,k})$ vertices in $T_{i,j,k}$ do not have semi-useful collision. For the case of $|T_{i,j,k}| \leq \sqrt{n}/\phi^{14}$, there exists a vertex $x \in T_{i,j,k}$ without a semi-useful collision, then by Lemma 6.4.21, Subroutine `Testing-Subset-Collision` rejects with probability at least $1 - \delta/2n$.

For the case of $|T_{i,j,k}| > \sqrt{n}/\phi^{14}$, by Lemma 6.4.17, if $|A_s(T_{i,j,k})| < (1 - 4\varepsilon_1)|T_{i,j,k}|\gamma_{i,j,k}$, then Subroutine `Testing-Subset-Collision` rejects with probability $1 - \delta$.

We now consider the case of $|A_s(T_{i,j,k})| \geq (1 - 4\varepsilon_1)|T_{i,j,k}|\gamma_{i,j,k}$. For a vertex $x \in T_{i,j,k}$ with $|W_x| = 1$, the number of semi-useful collision to x is at most $(1 + 3\varepsilon_1)(1 + \varepsilon_1)\gamma_{i,j,k}$ by Fact 6.4.4. Since a total weight of at least $10\varepsilon_1 \mathsf{weight}_w(T_{i,j,k})$ vertices in $T_{i,j,k}$ have no semi-useful collision, at least a $10(1-o(1))\varepsilon_1$ fraction of all the vertices in $T_{i,j}$ have no semi-useful collision, and then at least $(1-o(1))\varepsilon_1\gamma_{i,j,k}|T_{i,j,k}|$ semi-useful collisions are between vertices $L_{i,j,k}$ and $C_{i,j,k}$. Notice that every semi-useful collision is a good/intermediate collision,

$|Y| \geq (1 - o(1))\varepsilon_1\gamma_{i,j,k}|T_{i,j,k}|$. By Lemma 6.4.18, there exists a distort subset in $T_{i,j,k}$ or $C_{i,j,k}$. Then Step 2 of Subroutine `Testing-Subset-Collision` reject with probability at least $1 - \delta$. □

We prove some additional property for the subroutine `Testing-Subset-Collision`.

Let

$$R^0_{i,j,k} = \{y \in C_{i,j,k} : \exists v, w \in T_{i,j,k} - L_{i,j,k} \text{ s.t. } v \neq w, \\ y \text{ is a good/intermediate collision of both } v \text{ and } w\},$$

$$L^0_{i,j,k} = \{x \in T_{i,j,k} - L_{i,j,k} : \exists y \in R^0_{i,j,k} \text{ s.t. } y \text{ is a good/intermediate collision of } x\},$$

and Y^0 be the bipartite graph between $L^0_{i,j,k}$ and $C_{i,j,k}$ such that $x \in L^0_{i,j,k}$ is adjacent to $y \in C_{i,j,k}$ if y is a good/intermediate collision of x.

Fact 6.4.23. *If the distance between every pair of vertices in $T_{i,j,k}$ is not distorted, then Y^0 is an empty graph.*

Lemma 6.4.24. *If $|T_{i,j,k}| > \sqrt{n}/\phi^{14}$ and $|Y^0| \geq 4\phi^3|T_{i,j,k}|\gamma_{i,j,k}$, then there exists a distorted set in $T_{i,j,k}$.*

Proof. For any vertex $x \in L^0_{i,j,k}$, let

$$D_{Y^0}(x) = \{v \in L^0_{i,j,k} - \{x\} : \exists y \in R^0_{i,j,k} \text{ s.t. } y \text{ is adjacent to both } x \text{ and } v \text{ in } Y^0\}.$$

If there exists a vertex $x \in L^0_{i,j,k}$ such that $|D_{Y^0}(x)| \geq \frac{\phi^4|T_{i,j,k}|}{\sqrt{n}} - 1$, then $\{x\} \cup D_{Y^0}(x)$ is a distorted set for $T_{i,j,k}$, since every pair of vertices in $\{x\} \cup D_{Y^0}(x)$ has distance at most $4r$ in $\mathcal{M}$.

Now assume $|D_{Y^0}(x)| < \frac{\phi^4|T_{i,j,k}|}{\sqrt{n}} - 1$ for every $x \in L^0_{i,j,k}$. Since $L^0_{i,j,k} \subseteq T_{i,j,k} - L_{i,j,k}$, every vertex in $L^0_{i,j,k}$ has at most $(1 + 3\varepsilon_1)(1 + \varepsilon_1)\gamma_{i,j,k}$ neighbors in Y^0 by Fact 6.4.2. Thus, the number of vertices in $|L^0_{i,j,k}| \geq \frac{|Y^0|}{(1+3\varepsilon_1)(1+\varepsilon_1)\gamma_{i,j,k}} \geq 2\phi^3|T_{i,j,k}|$, and then we have $1 \geq \frac{2\phi^3|T_{i,j,k}|}{|L^0_{i,j,k}|} \geq \frac{2\phi^3|T_{i,j,k}|^2}{|L^0_{i,j,k}|n}$. Since for every vertex x in $L^0_{i,j,k}$, there exists a $v \in L^0_{i,j,k}$ such that the distance between x and v is distorted, $L^0_{i,j,k}$ is a distorted set in $T_{i,j,k}$. □

Let $U_{i,j,k} = \{v \in T_{i,j,k} - L_{i,j,k} - L^0_{i,j,k} : |(v, A_s(v)) - (Y \cup Y^0)| \geq (1 - 4\varepsilon_1)\gamma_{i,j,k}\}$, and $SU_{i,j,k} = \{v \in S_{i,j,k} : Assignto(v) \in U_{i,j,k}\}$.

Given a subset Q of $T_{i,j,k}$, define bipartite graph $A_b(Q) = (Q, C_{i,j,k})$ such that there is an edge between $x \in Q$ and $y \in C_{i,j,k}$ if and only if $x \in Q \cap (T_{i,j,k} - L_{i,j,k} - L^0_{i,j,k})$ and y is a good/intermediate collision to x, but not a semi-useful collision to x.

Lemma 6.4.25. *If $A_s(T_{i,j,k}) \geq (1-4\varepsilon_1)|T_{i,j,k}|\gamma_{i,j,k}$, $A_b(T_{i,j,k}) \leq 2\varepsilon_1^2|T_{i,j,k}|\gamma_{i,j,k}$, $|Y| \leq \phi|T_{i,j,k}|\gamma_{i,j,k}$ and $|Y^0| \leq 4\phi^3|T_{i,j,k}|\gamma_{i,j,k}$, then $|U_{i,j,k}| \geq (1-10\varepsilon_1)|T_{i,j,k}|$ and $|SU_{i,j,k}| \geq (1-12\varepsilon_1)|S_{i,j,k}|$.*

Proof. We have

$$|A_s(T_{i,j,k} - L_{i,j,k} - L^0_{i,j,k})| \geq |A_s(T_{i,j,k})| - |Y| - |Y^0| \geq (1 - 4(1+o(1))\varepsilon_1)|T_{i,j,k}|\gamma_{i,j,k}.$$

By the definition of Y and Y^0, every vertex in $T_{i,j,k} - L_{i,j,k} - L^0_{i,j,k}$ satisfying $|W_x| \leq 1$. By Fact 6.4.4, there are at least

$$\frac{(1-4(1+o(1))\varepsilon_1)|T_{i,j,k}|\gamma_{i,j,k}}{(1+3\varepsilon_1)(1+\varepsilon_1)\gamma_{i,j,k}} \geq (1-(1+o(1))8\varepsilon_1)|T_{i,j,k}|$$

vertices v in $T_{i,j,k} - L_{i,j,k} - L^0_{i,j,k}$ which have semi-useful collisions in $C_{i,j,k}$. Let $U'_{i,j,k}$ denote this set of vertices.

If a vertex in $T_{i,j,k} - L_{i,j,k} - L^0_{i,j,k}$ has a semi-useful collision, then this vertex has at least $(1-3\varepsilon_1)\gamma_{i,j,k}$ good/intermediate collisions. On the other hand, there are at most $\frac{|A_b(T_{i,j,k})|}{\varepsilon_1\gamma_{i,j,k}} \leq 2\varepsilon_1|T_{i,j,k}|$ vertices v in $U'_{i,j,k}$ satisfying $|A_b(v)| \geq \varepsilon_1\gamma_{i,j,k}$, and then there are at most $2\varepsilon_1|T_{i,j,k}|$ vertices in $U'_{i,j,k}$ but not in $U_{i,j,k}$. Hence $|U_{i,j,k}| \geq (1-10\varepsilon_1)|T_{i,j,k}|$.

Since for every vertex x in $T_{i,j,k}$, $\beta_{i,j,k} \leq |Assign(x)| < (1+\varepsilon_1)\beta_{i,j,k}$, $\beta_{i,j,k}|T_{i,j,k}| \leq |S_{i,j,k}| < (1+\varepsilon_1)\beta_{i,j,k}|T_{i,j,k}|$. Then $|SU_{i,j,k}| \geq \beta_{i,j,k}|U_{i,j,k}| \geq (1-12\varepsilon_1)|S_{i,j,k}|$. □

Lemma 6.4.26. *If $|T_{i,j,k}| > \sqrt{n}/\phi^{14}$, $A_b(T_{i,j,k}) \leq 2\varepsilon_1^2|T_{i,j,k}|\gamma_{i,j,k}$, $A_s(T_{i,j,k}) \geq (1-4\varepsilon_1)|T_{i,j,k}|\cdot\gamma_{i,j,k}$, $|Y| \leq \phi|T_{i,j,k}|\gamma_{i,j,k}$ and $|Y^0| \leq 4\phi^3|T_{i,j,k}|\gamma_{i,j,k}$, then with probability $1 - O(\phi^{16})$, the set Q in Subroutine* `Testing-Collision-`$T_{i,j,k}$ *satisfies*

1. *All the sampled vertices in Q are distinct;*
2. *$|Q \cap U_{i,j,k}| \geq (1-12\varepsilon_1)|Q|$;*

Proof. By Lemma 6.4.16, with probability $1 - O(\phi^{16})$ all the sampled vertices are distinct.

By Lemma 6.4.25, the probability that one vertex sample in 5(a) is in $U_{i,j,k}$ is at least $1 - 10\varepsilon_1$. By Chernoff bound,

$$\Pr[|Q \cap U_{i,j,k}| \geq (1 - 12\varepsilon_1)|Q|] \geq 1 - \exp(-\Omega(\varepsilon_1^2|Q|)).$$

By union bound, we obtain the lemma. □

Definition 6.4.27. *Let (v, y) be a pair of vertices satisfying $v \in T_{i,j,k}$ and y is a collision of v. We say (v, y) is nice if*

1. *v is in $T_{i,j,k} - L_{i,j,k} - L^0_{i,j,k}$.*
2. *y is a semi-useful collision of v.*

Let y be a vertex in $S_{i,j,k}$, we say a pair (v, y) is nice if $(Assignto(v), y)$ is a nice pair.

Fact 6.4.28. *Given a set of vertices $T_{i,j,k}$, we have*

1. *If (v, y) is a nice pair for $T_{i,j,k}$, then y is not a good/intermediate collision for all the vertices in $T_{i,j,k}$ except v.*
2. *Fix a vertex $y \in C_{i,j,k}$, if (v, y) is a nice pair for some vertex $v \in T_{i,j,k}$, then y forms a nice pair with at least $\beta_{i,j,k}$ and at most $(1 + \varepsilon_1)\beta_{i,j,k}$ vertices in $S_{i,j,k}$.*

Lemma 6.4.29. *If the Q obtained in step 5(a) of Subroutine* `Tesing-Collision-`$T_{i,j,k}$ *satisfies*

1. *All the sampled vertices in Q are distinct,*
2. *Every pair of vertices in Q has distance at least $4r + 4\delta$ within $\mathcal{M}$*
3. *$|Q \cap U_{i,j,k}| \geq (1 - 12\varepsilon_1)|Q|$,*
4. *$|A_n(Q)| \leq \varepsilon_1^2\phi^{10}|Q|\gamma_{i,j,k}$,*
5. *$|(Q, C_{i,j,k}) \cap Y| \leq \phi|Q|\gamma_{i,j,k}/\varepsilon_2$*

then

1. *With probability at least $1 - 15\varepsilon_1$, Subroutine* `Testing-Collision-Subset` *returns a nice pair of vertices.*

2. *For any* $x \in (T_{i,j,k} - L_{i,j,k} - L^0_{i,j,k}) \cap Q$, *Subroutine* `Testing-Collision-Subset` *returns a nice pair containing* x *with probability at most* $\frac{(1+22\varepsilon_1)|T_{i,j,k}|}{|Q||S_{i,j,k}|}$.

3. *Fix a nice pair of vertices* (x, y) *with* $x \in S_{i,j,k}$ *and* $Assignto(x) \in (T_{i,j,k} - L_{i,j,k} - L^0_{i,j,k}) \cap Q$, *Subroutine* `Testing-Collision-Subset` *returns the pair* (x, y) *with probability at most* $\frac{(1+22\varepsilon_1)|T_{i,j,k}|}{\gamma_{i,j,k}|Q||S_{i,j,k}|}$.

Proof. Since for every pair of vertices in Q has distance at least $4r + 4\delta$, every vertex $y \in C_{i,j,k}$ have at most one $x \in Q$ satisfying $\mathcal{M}(x, y) \leq r$.

For any $v \in U_{i,j,k}$, there are at least $(1 - 4\varepsilon_1)\gamma_{i,j,k}$ nice pairs containing v by definition, and at most $(1 + 3\varepsilon_1)(1 + \varepsilon_1)\gamma_{i,j,k}$ nice pairs. Let X_i be the indicator variable that i-th sample of step 1 in Subroutine `Testing-Collision-Subset` forms a nice pair with a vertex in Q. We have

$$\frac{(1 - 4\varepsilon_1)\gamma_{i,j,k}|Q \cap U_{i,j,k}|}{|C_{i,j,k}|} \leq \Pr[X_i] \leq \frac{(1 + 3\varepsilon_1)(1 + \varepsilon_1)\gamma_{i,j,k}|Q|}{|C_{i,j,k}|}.$$

So the expected number of nice pairs is

$$\begin{aligned}\frac{(1 - 4\varepsilon_1)(1 - 12\varepsilon_1)\gamma_{i,j,k}|Q||K|}{|C_{i,j,k}|} \leq & \frac{(1 - 4\varepsilon_1)\gamma_{i,j,k}|Q \cap U_{i,j,k}||K|}{|C_{i,j,k}|} \\ \leq & \mathbb{E}[X = \sum X_i] \\ \leq & \frac{(1 + 3\varepsilon_1)(1 + \varepsilon_1)\gamma_{i,j,k}|Q||K|}{|C_{i,j,k}|}.\end{aligned}$$

By Corollary 6.3.5, $\frac{1-16\varepsilon_1}{2\phi^{11}} \leq \mathbb{E}[X]$. By Chernoff bound,

$$\Pr[\frac{(1 - 17\varepsilon_1)\gamma_{i,j,k}|Q||K|}{|C_{i,j,k}|} \leq X \leq \frac{(1 + 5\varepsilon_1)\gamma_{i,j,k}|Q||K|}{|C_{i,j,k}|}] \geq 1 - \exp(-1/\phi^{10})$$

On the other hand, let Z_i be the indicator variable that i-th sampled vertex z of step 1 in Subroutine `Testing-Collision-Subset` satisfying that there is a vertex $x \in Q$ with $\mathcal{M}(x, z) \leq r$, but (x, z) is not a nice pair. We have

$$\Pr[Z_i] = \frac{|A_n(Q)| + |(Q, C_{i,j,k} \cap Y)| + 12\varepsilon_1|Q|(1 + 3\varepsilon_1)(1 + \varepsilon_1)\gamma_{i,j,k}}{|C_{i,j,k}|} \leq \frac{13\varepsilon_1\gamma_{i,j,k}|Q|}{|C_{i,j,k}|}.$$

So, $\mathbb{E}[Z = \sum Z_i] \leq \frac{13\varepsilon_1\gamma_{i,j,k}|Q||K|}{|C_{i,j,k}|}$, and again, by Chernoff bound,

$$\Pr[Z \geq \frac{14\varepsilon_1\gamma_{i,j,k}|Q||K|}{|C_{i,j,k}|}] \leq \exp(-1/\phi).$$

With probability $1-\exp(-1/\phi)$, the total number of nice pairs is at least $\frac{1-17\varepsilon_1}{14\varepsilon_1}$ times of the total number of non-nice pairs. Thus, the overall probability of obtaining a nice pair is at least $1-15\varepsilon_1$.

Fix a vertex $x \in S_{i,j,k}$ with $v = Assignto(x) \in (T_{i,j,k}-L_{i,j,k}-L^0_{i,j,k})\cap Q$, let p_v denote the probability that a sample in K forms a nice pair with v. So p_v is at most $\frac{(1+3\varepsilon_1)(1+\varepsilon_1)\gamma_{i,j,k}}{|C_{i,j,k}|}$. For a non-negative integer α, the probability that there are α samples forming nice pairs with v is $\binom{|K|}{\alpha}(1-p_v)^{|K|-\alpha}p_v^\alpha$, and the probability that the returned pair containing v is $\alpha/(X+Z)$, where $X+Z$ is within $[\frac{(1-17\varepsilon_1)\gamma_{i,j,k}|Q||K|}{|C_{i,j,k}|}, \frac{(1+20\varepsilon_1)\gamma_{i,j,k}|Q||K|}{|C_{i,j,k}|}]$ with probability at least $1-\exp(-1/\phi)$. Let q_v be the overall probability that the returned pair is a nice pair containing x such that $Assign(x) = v$. We have

$$\begin{aligned} q_v \leq & (1-\exp(-1/\phi))\sum_{\alpha=0}^{|K|}\binom{|K|}{\alpha}(1-p_v)^{|K|-\alpha}p_v^\alpha\frac{\alpha}{(1-17\varepsilon_1)\gamma_{i,j,k}|Q||K|/|C_{i,j,k}|}+\exp(-1/\phi) \\ \leq & \frac{1+(1+o(1))21\varepsilon_1}{|Q|}. \end{aligned}$$

Since every vertex v in $T_{i,j,k}$ have at least $\beta_{i,j,k}$ and at most $(1+\varepsilon_1)\beta_{i,j,k}$ vertices assigned to v, and $\beta_{i,j,k}|T_{i,j,k}| \leq |S_{i,j,k}| \leq (1+\varepsilon_1)\beta_{i,j,k}|T_{i,j,k}|$, so the probability of returning a nice pair containing x is at most $\frac{1+(1+o(1))21\varepsilon_1}{|Q|(1+\varepsilon_1)^j} \leq \frac{(1+22\varepsilon_1)|T_{i,j,k}|}{|Q||S_{i,j,k}|}$.

Fix a nice pair (x,y) with $v = Assignto(x) \in (T_{i,j,k}-L_{i,j,k}-L^0_{i,j,k})\cap Q$, the probability of sampling y is $1/|C_{i,j,k}|$. For an non-negative integer β, the probability that there are β samples hitting y is $\binom{|K|}{\beta}(1-1/|C_{i,j,k}|)^{|K|-\beta}(1/|C_{i,j,k}|)^\beta$, and the probability of choosing (v,y) is $\beta/(X+Z)$. Thus, the overall probability of choosing (v,y) is at most

$$\begin{aligned} & \left(1-\exp\left(-\frac{1}{\phi}\right)\right)\sum_{\beta=0}^{|K|}\binom{|K|}{\beta}\left(1-\frac{1}{|C_{i,j,k}|}\right)^{|K|-\beta}\left(\frac{1}{|C_{i,j,k}|}\right)^\beta\frac{\beta|C_{i,j,k}|}{(1-17\varepsilon_1)\gamma_{i,j,k}|Q||K|} \\ & +\exp\left(-\frac{1}{\phi}\right) \\ \leq & \frac{1+21\varepsilon_1}{\gamma_{i,j,k}|Q|}. \end{aligned}$$

Hence, the probability of returning (x,y) is at most $\frac{1+21\varepsilon_1}{\gamma_{i,j,k}|Q|(1+\varepsilon_1)^j} \leq \frac{(1+22\varepsilon_1)|T_{i,j,k}|}{\gamma_{i,j,k}|Q||S_{i,j,k}|}$. □

Lemma 6.4.30. *For any $T_{i,j,k}$ with $|T_{i,j,k}| > \sqrt{n}/\phi^{14}$ and $\mathsf{weight}(T_{i,j,k}) \geq \varepsilon_2^2 n$, if one of following conditions hold*

1. $A_s(T_{i,j,k}) < (1-4\varepsilon_1)|T_{i,j,k}|\gamma_{i,j,k}$
2. $A_b(T_{i,j,k}) \geq 2\varepsilon_1^2|T_{i,j,k}|\gamma_{i,j,k}$
3. $|Y| \geq \phi|T_{i,j,k}|\gamma_{i,j,k}$
4. $|Y^0| \geq 4\phi^3|T_{i,j,k}|\gamma_{i,j,k}$

then Subroutine `Testing-Collision-`$T_{i,j,k}$ *rejects with probability at least* $1-O(\delta/n)$.

Proof. By Lemma 6.4.18 and Lemma 6.4.24, if the third or fourth conditions hold, then the subroutine rejects with probability at least $1-\delta/n$.

Now we consider the case of $A_s(T_{i,j,k}) < (1-4\varepsilon_1)|T_{i,j,k}|\gamma_{i,j,k}$. Let X_i be the random variable of $A_s(v)$ for i-th sampled vertex v. $\mathbb{E}[X = \sum X_i] \leq |A_s(T_{i,j,k})||Q|/|T_{i,j,k}| \leq (1-4\varepsilon_1)\gamma_{i,j,k}|Q|$. By Markov inequality, $\Pr[X \geq (1-2\varepsilon_1)\gamma_{i,j,k}|Q|] \leq \frac{1-4\varepsilon_1}{1-2\varepsilon_1}$. With probability at least ε_1, a random set Q satisfies $A_s(Q) < (1-2\varepsilon_1)|Q|\gamma_{i,j,k}$. Hence, with probability $1-\exp(-\log(4/\delta)/\phi^7)$, there are $\Omega(\log(4/\delta)/\phi^7)$ samples of Q satisfying $A_s(Q) < (1-2\varepsilon_1)|Q|\gamma_{i,j,k}$. By Lemma 6.4.15, with overall probability at least $1-O(\delta/n)$, the subroutine rejects.

Consider the case of $A_b(T_{i,j,k}) \geq 2\varepsilon_1^2|T_{i,j,k}|\gamma_{i,j,k}$. Let Z_i be the random variable of $A_s(v)$ for i-th sampled vertex v. We have $0 \leq Z_i \leq (1+3\varepsilon_1)(1+\varepsilon_1)\gamma_{i,j,k}$, and $\mathbb{E}[Z = \sum Z_i] = A_b(T_{i,j,k})|Q|/|T_{i,j,k}| \geq 2\varepsilon_1^2\gamma_{i,j,k}|Q|$. By Hoeffding bound, we have $\Pr[Z \geq \varepsilon_1^2\gamma_{i,j,k}|Q|] \geq 1-\exp(-\Omega(\varepsilon_1^2|Q|))$. By Lemma 6.4.16, with probability at least $1-O(\phi^{16})$, all the sampled vertices in Q are distinct. Hence, with probability $1-O(\phi^{16})$, a random set of Q satisfying $A_b(Q) \geq \varepsilon_1^2\gamma_{i,j,k}|Q|$. By Lemma 6.4.15, the subroutine rejects such a set with probability at least $1-O(\delta/n)$. □

Lemma 6.4.31. *For any* $T_{i,j,k}$ *with* $|T_{i,j,k}| > \sqrt{n}/\phi^{14}$ *and* $\mathsf{weight}(T_{i,j,k}) \geq \varepsilon_2^2 n$, *either Subroutine* `Testing-Collision-`$T_{i,j,k}$ *rejects with probability at least* $1-O(\delta/n)$ *or Subroutine* `Testing-Collision-`$T_{i,j,k}$ *satisfies following conditions*

1. *With probability at least* $1-17\varepsilon_1$, *the subroutine returns a nice pair of collision.*
2. *Fix a vertex* $x \in S_{i,j,k}$ *such that* $Assignto(x) \in U_{i,j,k}$, *the probability that the subroutine returns a nice pair of collision containing* x *is at most* $\frac{1+22\varepsilon_1}{|S_{i,j,k}|}$.

3. *Fix a pair of nice collision (x, y) with $Assignto(x) \in U_{i,j,k}$, the algorithm returns the pair (x, y) with probability at most $\frac{1+22\varepsilon_1}{\gamma_{i,j,k}|S_{i,j,k}|}$.*

Proof. We first consider the case of $|T_{i,j,k}| > \sqrt{n}/\phi^{14}$. By Lemma 6.4.30, we assume $A_s(T_{i,j,k}) \geq (1-4\varepsilon_1)|T_{i,j,k}|\gamma_{i,j,k}$, $A_b(T_{i,j,k}) < 2\varepsilon_1^2|T_{i,j,k}|\gamma_{i,j,k}$, $|Y| < \phi|T_{i,j,k}|\gamma_{i,j,k}$ and $|Y^0| < 4\phi^3|T_{i,j,k}|\gamma_{i,j,k}$.

Let p be the probability that at least one of following two conditions satisfy for a random set of Q

1. There exists $v, w \in Q$ such that the distance between v and w is distorted.
2. $|A_n(Q)| > \varepsilon_1^2\phi^{10}|Q|\gamma_{i,j,k}$.

The probability that none of the samples of Q in step 5(a) satisfying at least one of the above two conditions is $(1-p)^{\log(4/\delta)/\phi^8}$. If $p > \phi^8$, then with probability at least $1-\delta/4$, there is a sample of Q satisfying at least one of the above two conditions. By Lemma 6.4.15, the subroutine rejects with probability at least $1-\delta/n$.

Now we assume $p < \phi^8$. By Lemma 6.4.16, with probability $1-O(\phi^{16})$, all the sampled vertices in one run of step 5(a) are distinct. By Lemma 6.4.25, the probability that one vertex sample in 5(a) is in $U_{i,j,k}$ is at least $(1-10\varepsilon_1)$. By Chernoff bound,

$$\Pr[|Q \cap U_{i,j,k}| \geq (1-12\varepsilon_1)|Q|] \geq 1 - \exp(-\Omega(\varepsilon_1^2|Q|)).$$

By Markov inequality, we have $|(Q, C_{i,j,k}) \cap Y| > \phi|Q|\gamma_{i,j,k}/\varepsilon_2$ with probability $1-\varepsilon_2$. Hence, the probability that a random set of Q satisfying all the five conditions of Lemma 6.4.29 is at least $1-2\varepsilon_2$, and by Lemma 6.4.29, the probability of returning a nice pair is at least $1-15\varepsilon_1$. With probability at most $1-\exp(-1/\phi)$, a fraction of at most $16\varepsilon_1$ returned pairs are not nice pairs. Hence, with overall probability at least $1-17\varepsilon_1$, the first condition holds.

Let q_v for $v \in U_{i,j,k}$ denote the probability that a set Q sampled in step 5(a) of subroutine `Testing-Collision-`$T_{i,j,k}$ satisfying the $v \in Q$. For any $u, v \in U_{i,j,k}$, $q_u = q_v$. Hence $(1-O(1/\sqrt{n}))\frac{|Q|}{|T_{i,j,k}|} \leq q_v \leq \frac{|Q|}{|T_{i,j,k}|}$. By Lemma 6.4.29, for a random set of Q, the probability of returning $x \in S_{i,j,k}$ with $Assignto(x) \in U_{i,j,k}$ is at most $\frac{1+22\varepsilon_1}{|S_{i,j,k}|}$.

Similarly, by Lemma 6.4.29, the probability of returning (x, y) with $Assignto(x) \in U_{i,j,k}$ is at most $\frac{1+22\varepsilon_1}{\gamma_{i,j,k}|S_{i,j,k}|}$.

Now we consider the case of $|T_{i,j,k}| < \sqrt{n}/\phi^{14}$. By Lemma 6.4.21 and Lemma 6.4.14, the first condition holds. The probability that returning a pair containing $x \in S_{i,j,k}$ with $v = Assignto(x)$ is at most $\frac{1}{|T_{i,j,k}||Assignto(v)|} \leq \frac{1+\varepsilon_1}{|S_{i,j,k}|}$.

Fix an arbitrary nice pair (x, y) with $v = Assignto(x)$. If $|A_s(v) \cup A_n(v)|$ is smaller than $(1 - 2\varepsilon_1)\gamma_{i,j,k}|Q|/|C_{i,j,k}|$ or greater than $(1 + 5\varepsilon_1)\gamma_{i,j,k}|Q|/|C_{i,j,k}|$, then the algorithm rejects with probability at least $1 - \delta$ by Chernoff bound. Hence, with probability at least $1 - \delta/n$, there are at least $(1 - 3\varepsilon_1)\gamma_{i,j,k}|Q|/|C_{i,j,k}|$ vertices in K has distance at most r to v. The probability of returning (x, y) is at most

$$\frac{1}{|T_{i,j,k}|}\left(1 - \frac{\delta}{n}\right)\left(1 - (1 - \frac{1}{|C_{i,j,k}|})^{|Q|}\right)\frac{|C_{i,j,k}|}{(1 - 3\varepsilon_1)\gamma_{i,j,k}|Q|}\frac{1}{|Assignto(v)|} + \frac{\delta}{n} \leq \frac{1 + 22\varepsilon_1}{\gamma_{i,j,k}|S_{i,j,k}|}$$

□

Now we present an algorithm for the Testing Collision Problem.

Subroutine `Testing-Collision`**:**

Input: Two graphs $I, J \in \{G, H\}$. $T_{i,j,k}, S_{i,j,k} \subseteq V_I$ and $H_{i,j,k}, C_{i,j,k} \subseteq V_J$, $Assign(v)$ for every $v \in T_{i,j,k}$, $Assignto(v)$ for every $x \in S_{i,j,k}$, $\mathsf{weight}_w : \cup T_{i,j,k} \to \mathbb{R}^{\geq 0}$, and parameter δ.

1. For every $0 \leq i, j, k \leq 6 \log n/\varepsilon_1$ with $\mathsf{weight}_w(T_{i,j,k}) \geq \varepsilon_2^2 n$, run Subroutine `Testing-Subset-Collision` with $T_{i,j,k}, S_{i,j,k}, C_{i,j,k}$ and parameter $\varepsilon_2^2\delta$.
2. Reject if any execution of subroutine `Testing-Subset-Collision` rejects, otherwise accept.

Proof of Theorem 6.4.5. By Lemma 6.4.22, for any positive instance, the subroutine accepts with probability at least $\left(1 - \varepsilon_2^2\delta\right)^{1/\varepsilon_2^2} \geq 1 - \delta$.

For any negative instance, there are $0 \leq i, j, k \leq 6 \log n/\varepsilon_1$ such that a total weight of at least $10\varepsilon_1\mathsf{weight}_w(T_{i,j,k})$ vertices in $T_{i,j,k}$ do not have semi-useful collision. By Lemma 6.4.22, the subroutine rejects with probability at least $1 - \delta$.

□

6.5 Testing label bijection

In this section, we present an efficient algorithm solving the Testing Label Bijection problem, which bypasses the $\Omega(n^{2/3})$ lower bound of testing identity of two unknown distributions based on the estimation of neighbor distance metric. We first define the Testing Label Bijection problem.

Assuming $I, J \in \{G, H\}$ are two distinct graphs.

Definition 6.5.1. *A vertex $x \in S_{i,j,k}$ is semi-mathched by a vertex $y \in V_J$ through $z \in V_J$ if*

1. *z is a semi-useful collision of $v = Assignto(x) \in T_{i,j,k}$.*
2. *$\rho_3(x, y) \leq 2\sigma$*
3. *$\mathcal{M}_J(z, y) \leq 5r + 10\sigma$.*

Definition 6.5.2. *A vertex $x \in S_{i,j,k}$ is matched by $y \in V_J$ within distance ζ if*

1. *y is at most ϕ^2-distorted;*
2. *$\rho_2(x, y) \leq 4\sigma$ and $\mathcal{M}(x, y) \leq \zeta$;*
3. *Let $v = Assignto(x)$. There exists a vertex $z \in V_J$ such that z is a semi-useful collision of v satisfying $\mathcal{M}_J(y, z) \leq \mathcal{M}_I(x, v) + 2\delta + \zeta$.*

In this section, we solve the following problem.

Problem Testing Label Bijection
Input: Two graphs G and H
Output:

1. Accept with probability $1 - \delta$ if all of the following conditions hold

(a) Every pair of vertices $v, w \in V_G$ satisfies $\mathcal{M}_I(v,w) - 2\sigma \leq \mathcal{M}(v,w) \leq \mathcal{M}_I(v,w) + 2\sigma$.

(b) Every pair of vertices $y, z \in V_H$ satisfies $\mathcal{M}_J(y,z) - 2\sigma \leq \mathcal{M}(y,z) \leq \mathcal{M}_J(y,z) + 2\sigma$.

(c) There is a bijection $\pi : V_G \to V_H$ s.t. for any $v \in V_I$, $\mathcal{M}(v, \pi(v)) = 0$.

2. Reject with probability $1-\delta$ if for any mapping $\pi : V_G \to V_H$, at least $4\varepsilon n$ vertices x in A_G do not satisfy at least one of the following conditions

(a) $\mathcal{M}(x, \pi(x)) \leq 1200r \log n/\varepsilon_1$

(b) There is no vertex $y \in V_H$ such that y matches x within distance $40r$, and $\mathcal{M}_H(\pi(x), y) \leq 1200r \log n/\varepsilon_1$.

We show that

Theorem 6.5.3. *There is an algorithm solving the Testing Label Bijection Problem with probability at least* $1-\delta$ *using* $O((\sqrt{n}\log n) \cdot \frac{1}{\phi^{O(1)}} \cdot \log(1/\delta) \cdot \frac{1}{\varepsilon_2^2})$ *label queries with running time poly*$(n, \frac{1}{\phi}, \frac{1}{\varepsilon_2^2}, \log(1/\delta))$.

We prove Theorem 6.5.3 in the rest of this section. In Section 6.5.1, we present an algorithm to solve the Testing Vertex Matching problem, and in Section 6.5.3, we show how to use the algorithm for Testing Vertex Matching problem solving the Testing Label Bijection problem.

6.5.1 Testing vertex matching

We consider the following problem.

Problem Testing vertex Matching

Input: Two graphs I and J, vertex subsets $S_{i,j,k}, T_{i,j,k} \subseteq V_I, C_{i,j,k}, H_{i,j,k} \subseteq V_J$ for $0 \leq i,j,k \leq 6\log n/\varepsilon_1$, robust weight function $\mathsf{weight}_w : \cup T_{i,j,k} \to \mathbb{R}^{\geq 0}$ satisfying $\sum_{x \in \cup T_{i,j,k}} \geq (1-\varepsilon)n$, label query oracle $\mathcal{O}$, parameter δ.

Output:

1. Accept with probability $1-\delta$ if all of the following conditions hold
   (a) Every pair of vertices $v, w \in V_I$ satisfies $\mathcal{M}_I(v,w) - 2\sigma \leq \mathcal{M}(v,w) \leq \mathcal{M}_I(v,w) + 2\sigma$.
   (b) Every pair of vertices $y, z \in V_J$ satisfies $\mathcal{M}_J(y,z) - 2\sigma \leq \mathcal{M}(y,z) \leq \mathcal{M}_J(y,z) + 2\sigma$.
   (c) There is a bijection $f : V_I \to V_J$ s.t. for any $v \in V_I$, $\mathcal{M}(v, f(v)) = 0$.
2. Reject with probability $1-\delta$ if a total weight of at least $12\varepsilon_1 n$ vertices in S_I are not matched within distance $40r$.

We prove the following theorem in the rest of Section 6.5.1 by showing that Testing Vertex Matching Problem can be reduced to the Testing Collision problem, and thus Subroutine `Testing-Collision` solves the Testing Vertex Matching Problem.

Lemma 6.5.4. *There is an algorithm solving the Testing Vertex Matching Problem with probability at least $1-\delta$ using $O((\sqrt{n}\log n)\cdot\frac{1}{\phi^{O(1)}}\cdot\log(1/\delta)\cdot\frac{1}{\varepsilon_2^2})$ label queries with running time poly$(n, \frac{1}{\phi}, \frac{1}{\varepsilon_2^2}, \log(1/\delta))$.*

We start from a few more definitions and facts.

Definition 6.5.5. *A vertex $x \in S_{i,j,k}$ is first type false semi-matched by y through z if the distance between x and Assignto(x) is distorted by f_I.*

A vertex $x \in S_{i,j,k}$ is second type false semi-matched by y through z if x is not first type false semi-matched by y through z, and there does not exists a vertex $y' \in V_J$ satisfying that $\rho_2(x,y') \leq 4\delta$, $\mathcal{M}(x,y') \leq 40r$, $\mathcal{M}_J(y',z) \leq 40r$ and y' is at most ϕ^2-distorted.

A vertex $x \in S_I$ is true semi-matched by y through z if x is neither first type nor second type false semi-matched by y through z.

We have following observations for collision, semi-matching, and matching.

Fact 6.5.6. *For any $0 \leq i, j, k \leq 6\log n/\varepsilon_1$,*

1. *For any vertex $v \in T_{i,j,k}$, if $z \in V_J$ is a collision of v, then $z \in C_{i,j,k}$.*

2. *For any vertex $x \in S_{i,j,k}$, if $z \in C_{i,j,k}$ is a semi-useful collision of $Assignto(x)$, then x is semi-matched by a vertex $u \in H_{i,j,k}$ through z satisfying $\mathcal{M}_J(u,z) \leq \mathcal{M}_I(x, Assignto(x)) + 2\sigma$.*

3. *For any vertex $x \in S_{i,j,k}$, if x is matched by y within distance $40r$ or semi-matched by $y \in V_J$, then $y \in H_{i,j,k}$.*

4. *If a vertex $x \in S_{i,j,k}$ is true semi-matched by y through z, then there exists a vertex $y' \in V_J$ matching x within distance $40r$.*

Proof. For the first argument, if z is a collision of v, then there is a vertex $u \in Y_J$ satisfying $\rho_4(v,u) \leq 2\delta$ and $\mathcal{M}_J(z,u) \leq r + 2\delta$. Since w is a robust weight function, $u \in H_{i,j,k}$ and thus $z \in C_{i,j,k}$.

The second argument follows from the definition of $H_{i,j,k}$ and the robust weight function.

Now we prove the third argument. Using the first argument, there is a vertex $y \in C_{i,j,k}$ satisfying $\mathcal{M}_J(y,z) \leq r + 2\delta$ and $\rho_4(Assignto(x), y) \leq 2\sigma$. Hence there is a bijection $g : V_I \to V_J$ such that $g(Assignto(x)) = y$ and for any $v \in V_G$

$$\max\{|\mathcal{M}_I(Assignto(x), v) - \mathcal{M}_J(y, g(v))|, \rho_3(v, g(v))\} \leq 2\sigma.$$

Hence $\rho_3(x, g(x)) \leq 2\sigma$ and $\mathcal{M}_J(y, g(x)) \leq \mathcal{M}_I(Assignto(x), x) + 2\sigma \leq 4r + 8\sigma$. Since $\mathcal{M}_J(z, g(x)) \leq \mathcal{M}_J(z,y) + \mathcal{M}_J(y, g(x)) \leq 5r + 10\sigma$, x is semi-matched by $g(x)$.

For the fourth argument, there exists a vertex $y' \in V_J$ satisfying that $\rho_2(x,y') \leq 4\delta$, $\mathcal{M}(x,y') \leq 40r$, $\mathcal{M}_J(y',z) \leq 40r$ and y' is at most ϕ^2-distorted. By Definition 6.5.2, y matches x within distance $40r$. □

Lemma 6.5.7. *Fix a vertex $y \in H_{i,j,k}$, if there is a vertex $x \in S_{i,j,k}$ such that x is second type false semi-matched vertex by y, then vertex y is an at least ϕ^2-distorted vertex.*

Proof. Without loss of generality, assume x is second type false semi-matched by y through $z \in C_{i,j,k}$. Let $v = Assignto(x)$. Since $\rho_3(x,y) \leq 2\sigma$ and $\mathcal{M}_J(z,y) \leq 5r + 10\sigma$, either y is at least ϕ^2-distorted, or $\mathcal{M}(x,y) > 40r$.

In the following, we show that $\mathcal{M}(x,y) > 40r$ implies that y is at least ϕ^2-distorted. Assume $\mathcal{M}(x,y) > 40r$. We have

$$\mathcal{M}(x,y) \leq \mathcal{M}(x,v) + \mathcal{M}(v,z) + \mathcal{M}(z,y).$$

Since x is not first type semi-matched by y through z, $\mathcal{M}(x,v) \leq \mathcal{M}_I(x,v) + 2\sigma \leq 4r + 8\sigma$. Together with $\mathcal{M}(v,z) \leq r$, we have $\mathcal{M}(y,z) > 40r - 5r - 8\sigma = 35r - 8\sigma$. However, by Definition 6.5.1, $\mathcal{M}_J(y,z) \leq 5r + 10\sigma$.

Since z is a semi-useful collision to v, there are at least $(1-2\varepsilon_1)|B_I(v,r)|$ vertices $u \in B_J(z, 2r+6\sigma)$ satisfying $\mathcal{M}(v,u) \leq r$. For every such vertex u, we have $\mathcal{M}_J(u,y) \leq \mathcal{M}_J(u,z) + \mathcal{M}_J(z,y) \leq 2r + 6\sigma + 5r + 10\sigma = 7r + 16\sigma$. On the other hand,

$$\mathcal{M}(u,y) \geq \mathcal{M}(y,z) - \mathcal{M}(u,z) \geq \mathcal{M}(y,z) - (\mathcal{M}(u,v) + \mathcal{M}(v,z)) \geq 35r - 8\sigma - 2r \geq 33r - 8\sigma.$$

So, the distance between u and y is distorted. Thus, there are at least $(1-2\varepsilon_1)|B_I(v,r)|$ vertices in $B_J(z, 2r+6\sigma)$ have distance distorted to y.

On the other hand, $B_I(v,r) \subseteq B_I(x, 5r+6\sigma)$. Since $|B_I(x,40r)|/|B_I(x,r)| \leq 1/2\phi$, with $\rho_3(x,y) \leq 2\delta$, we have

$$|B_I(v,r)| \geq 2\phi |B_I(x, 5r+6\sigma)| \geq 2\phi |B_J(y, 5r+4\sigma)| \geq 4\phi^2 |B_J(y, 40r)|.$$

Using $B_J(z, 2r+6\sigma) \subseteq B_J(y, 7r+16\sigma)$, vertex y is at least ϕ^2 distorted. □

Lemma 6.5.8. *Let $y \in H_{i,j,k}$ be an at least ϕ^2-distorted vertex. If there are totally at most $\phi^4 |H_{i,j,k}|/\sqrt{n}$ vertices in $H_{i,j,k}$ at least ϕ^2-distorted, then there are at most $16\phi^3 |H_{i,j,k}|/\sqrt{n}$ vertices in $S_{i,j,k}$ second type false semi-matched by y.*

Proof. Fix a vertex $z \in C_{i,j,k}$ satisfying that there is a vertex $x \in S_{i,j,k}$ semi-matched by y through z, let

$$D_z = \{u \in C_{i,j,k} : \mathcal{M}_J(y,u) \leq 7r + 16\sigma \text{ and } \mathcal{M}(z,u) \leq 2r\}$$

and

$$C_z = \{u \in C_{i,j,k} : \mathcal{M}_J(y,u) \leq 7r + 16\sigma \text{ and } \mathcal{M}(z,u) \leq 4r + 2\sigma\}.$$

Since x is second type false semi-matched by y through z, z is a semi-useful collision to $v = Assignto(x)$. Let v' be the vertex in $H_{i,j,k}$ satisfying $\rho_4(v',v) \leq 2\sigma$ and $\mathcal{M}_J(v',z) \leq$

$r + 2\sigma$. We have $\mathcal{M}_J(v', y) \leq \mathcal{M}_J(v', z) + \mathcal{M}_J(z, y) \leq r + 2\sigma + 5r + 10\sigma \leq 6r + 12\sigma$. Hence, at least $(1 - 2\varepsilon_1)|B_I(v, r)|$ vertices $u \in B_J(z, 2r + 6\sigma)$ satisfying $\mathcal{M}(v, u) \leq r$. Since $\mathcal{M}_J(u, y) \leq \mathcal{M}_J(u, z) + \mathcal{M}_J(z, y) \leq 2r + 6\sigma + 5r + 10\sigma \leq 7r + 16\sigma$, these vertices u are in D_z, and thus

$$\begin{aligned}
|C_z| &\geq |D_z| \\
&\geq (1 - 2\varepsilon_1)|B_I(v, r)| \\
&\geq 2(1 - 2\varepsilon_1)\phi|B_I(v, 40r)| \\
&> \phi|B_J(v', 40r - 2\delta)| \\
&\geq \phi|B_J(y, 7r + 16\sigma)|.
\end{aligned}$$

Let G_z be the set of vertices in $S_{i,j,k}$ which are second type false semi-matched by y through some vertex in C_z. We prove that $|G_z|$ is at most $8\phi^4|H_{i,j,k}|/\sqrt{n}$ by contradiction. Assume $|G_z| > 8\phi^4|H_{i,j,k}|/\sqrt{n}$. For every vertex $x \in G_z$, assume y semi-matches x through $u \in C_z$, we have

$$\mathcal{M}(x, z) \leq \mathcal{M}(x, Assignto(x)) + \mathcal{M}(Assignto(x), u) + \mathcal{M}(u, z) \leq 4r + 8\sigma + r + 4r + 2\sigma \leq 9r + 10\sigma.$$

Hence, for every pair of vertices v, t in G_z satisfying

$$\mathcal{M}(v, t) \leq \mathcal{M}(v, z) + \mathcal{M}(z, t) \leq 18r + 20\sigma.$$

If for every vertex $x \in G_z$, there are at least $|G_z| - 4\phi^4|H_{i,j,k}|/\sqrt{n}$ vertices in G_z with distance distorted to x. Then G_z is a distorted set for $H_{i,j,k}$. Otherwise, there is a vertex $x \in G_z$ such that there are at least $4\phi^4|H_{i,j,k}|/\sqrt{n}$ vertices in G_z with distance not distorted to x. Let set U_x be this set of vertices. Since $\rho_3(x, y) \leq 2\sigma$, there exists a bijection $g : V_G \to V_H$ such that $g(x) = y$ and for any $v \in V_G$

$$\max\{|\mathcal{M}_I(x, v) - \mathcal{M}_J(y, g(v))|, \rho_2(v, g(v))\} \leq 2\sigma.$$

Thus, for any $v \in U_x$, $\rho_2(v, g(v)) \leq 2\sigma$. Since there are at most $\phi^4|H_{i,j,k}|/\sqrt{n}$ vertices that are at least ϕ^2-distorted, there is a $v \in U_x$ such that $g(v)$ is a vertex less than ϕ^2-distorted. Let $t \in C_z$ be the vertex such that v is second type false semi-matched by y through t. We have

$$\mathcal{M}_J(y, g(v)) \leq \mathcal{M}_I(x, v) + 2\sigma \leq \mathcal{M}(x, v) + 4\sigma \leq 18r + 24\sigma$$

(the last inequality uses the condition of $v \in U_x$), and

$$\begin{aligned}\mathcal{M}_J(t, g(v)) \leq & \mathcal{M}_J(t, z) + \mathcal{M}_J(z, y) + \mathcal{M}_J(y, g(v)) \\ \leq & 7r + 16\sigma + 5r + 10\sigma + \mathcal{M}_J(y, g(v)) \\ \leq & 30r + 50\sigma.\end{aligned}$$

Now we prove $\mathcal{M}(z, g(v)) \leq 28r$ by contradiction. Assume $\mathcal{M}(z, g(v)) > 28r$, then every vertex $u \in D_z$ satisfies $\mathcal{M}(g(v), u) \geq \mathcal{M}(g(v), z) - \mathcal{M}(z, u) > 28r - 2r = 26r$ and $\mathcal{M}_J(g(v), u) \leq \mathcal{M}_J(g(v), y) + \mathcal{M}_J(y, u) \leq 18r + 24\sigma + 7r + 16\sigma \leq 25r + 40\sigma$. So, at least

$$|D_z| \geq \phi |B_J(y, 7r + 16\sigma)| \geq 2\phi^2 |B_J(y, 40r)| \geq 2\phi^2 |B_J(g(v), 22r - 24\sigma)|$$

vertices in $B_J(y, 7r + 16\sigma)$ have distance distorted to $g(v)$. Thus, $g(v)$ is an at least ϕ^2-distorted vertex, a contradiction.

On the other hand, $\mathcal{M}(v, z) \leq \mathcal{M}(v, Assignto(v)) + \mathcal{M}(Assignto(v), w) + \mathcal{M}(w, z) \leq 5r + 12\delta + r + 4r + 2\sigma \leq 10r + 14\sigma$, and thus

$$\mathcal{M}(v, g(v)) \leq \mathcal{M}(v, z) + \mathcal{M}(z, g(v)) \leq 10r + 14\sigma + 28r \leq 38r + 14\sigma.$$

So, v is not second type false semi-matched by y through w, a contradiction. Hence $|G_z| \leq 8\phi^4 |H_{i,k}|/\sqrt{n}$.

We find a sequence of vertices $z_1, z_2, \ldots, z_k \in B_J(y, 5r + 10\sigma)$ such that for each z_i, there exists a $x \in S_{i,j,k}$ which is second type false semi-matched by y through z_i, and $z_i \notin \cup_{j<i} C_{z_j}$. We have $D_{z_i} \cap D_{z_j} = \emptyset$, then

$$k \leq \frac{|B_J(y, 7r + 16\sigma)|}{\min_i |D_{z_i}|} \leq \frac{|B_J(y, 7r + 16\sigma)|}{\phi |B_J(y, 7r + 16\sigma)|}.$$

Thus $k \leq 1/\phi$. Then the lemma follows. □

By Lemma 6.5.7 and Lemma 6.5.8, we have

Corollary 6.5.9. *If following two conditions are satisfied*

1. *there are totally at most $\phi^4 |H_{i,j,k}|/\sqrt{n}$ vertices in $H_{i,j,k}$ that are at least ϕ^2-distorted,*
2. *at most $|S_{i,j,k}|/\sqrt{n}$ vertices x in $S_{i,j,k}$ have the distance distorted to x,*

then there are at most $16\phi^7|H_{i,j,k}|^2/n + |S_{i,j,k}|/\sqrt{n}$ *vertices in* $S_{i,j,k}$ *semi-matched, but not true semi-matched.*

Lemma 6.5.10. *Let* $M \subseteq T_{i,j,k}$ *be the set of vertices with semi-useful collisions. If following two conditions are satisfied*

1. *there are totally at most* $\phi^4|H_{i,j,k}|/\sqrt{n}$ *vertices in* $H_{i,j,k}$ *that are at least* ϕ^2*-distorted,*
2. *at most* $|S_{i,j,k}|/\sqrt{n}$ *vertices* x *in* $S_{i,j,k}$ *have the distance distorted to* $Assignto(x)$*,*

then a total weight of at least $\mathsf{weight}_w(M) - 64\phi^7 n$ *vertices in* $S_{i,j,k}$ *are matched by vertices in* $C_{i,j,k}$*.*

Proof. By Fact 6.5.6, a total weight of $\mathsf{weight}_w(M)$ are semi-matched by some vertices in $H_{i,j,k}$. By Corollary 6.5.9, a total weight of at least

$$\mathsf{weight}_w(M) - (16\phi^7|H_{i,j,k}|^2/n + |S_{i,j,k}|/\sqrt{n})(1+\varepsilon_1)\alpha_{i,j,k}$$

vertices in $S_{i,k}$ are true semi-matched. By Fact 6.3.3, $|H_{i,j,k}| \leq \min\{n, \frac{n}{(1-6\varepsilon_1)\alpha_{i,j,k}}\}$, and thus

$$\frac{16\phi^7(1+\varepsilon_1)\alpha_{i,j,k}|H_{i,j,k}|^2}{n} \leq 32\phi^7 \min\{\alpha_{i,j,k}n, \frac{n}{\alpha_{i,j,k}}\} \leq 32\phi^7 n.$$

We also have $(1+\varepsilon_1)\alpha_{i,j,k}|S_{i,j,k}|/\sqrt{n} \leq 2\sqrt{n}$. Hence a total weight of at least

$$\mathsf{weight}_w(M) - 32\phi^7 n - 2\sqrt{n} \geq \mathsf{weight}_w(M) - 64\phi^7 n$$

vertices in $S_{i,k}$ are true semi-matched. By Fact 6.5.6, we obtain the lemma. □

Proof of Lemma 6.5.4. By Theorem 6.4.5, the subroutine accepts with probability at least $1-\delta$ for any accept instance.

Now we consider an negative instance. Since a total weight of at least $12\varepsilon_1 n$ vertices are not matched by vertices in V_J, there are $0 \leq i,j,k \leq 6\log n/\varepsilon_1$ with $\mathsf{weight}_w(T_{i,j,k}) \geq \varepsilon_2^2 n$ such that a total weight of at least $11\varepsilon_1\mathsf{weight}_w(T_{i,j,k})$ vertices in $S_{i,j,k}$ are not matched by vertices in V_J. By Lemma 6.4.22 and Lemma 6.5.10, the subroutine rejects with probability at least $1-\delta$. □

6.5.2 Flow index

Definition 6.5.11. *For any $I \in \{G, H\}$, and a vertex $y \in \mathcal{M}_I$, let*

$$\alpha_1(y,z) = \begin{cases} 1 & \text{if } \mathcal{M}_I(y,z) \le r - 3\mu \\ (1 - \frac{\varepsilon_1}{64(\log n)^2})^{(\mathcal{M}_I(y,z)-r+3\mu)/\sigma} & \text{if } \mathcal{M}_I(y,z) > r - 3\mu \end{cases}$$

and $\alpha_1(y) = \sum_z \alpha_1(y,z)$, let

$$\beta_1(y,z) = \begin{cases} 1 & \text{if } |\mathcal{M}_I(y,z) - r| \le 2\mu \\ (1 - \frac{\varepsilon_1}{64(\log n)^2})^{(|\mathcal{M}_I(y,z)-r|-2\mu)/\sigma} & \text{if } |\mathcal{M}_I(y,z) - r| > 2\mu \end{cases}$$

and $\beta_1(y) = \sum_z \beta_1(y,z)$, and let

$$t_1(y) = \begin{cases} 1 & \text{if } \beta_1(y)/\alpha_1(y) \le \varepsilon_1/3 \\ \left(\frac{\varepsilon_1 \alpha_1(y)}{3\beta_1(y)}\right)^{8\log n} & \text{if } \beta_1(y)/\alpha_1(y) > \varepsilon_1/3 \end{cases}$$

Lemma 6.5.12. *If $\mu \ge 128(\log n)^3\sigma/\varepsilon_1$, then for any vertex $y \in V_I$*

1. *If $|B_I(y, r+2\mu)|/|B_I(y, r-2\mu)| \ge 1 + \varepsilon_1/2$, then $t_1(y) \le 2/n^2$;*
2. *If $|B_I(y, r+3\mu)/|B_I(y, r-3\mu)| \le 1 + \varepsilon_1/3$, then $1 - o(1) \le t_1(y) \le 1$;*
3. *If $\rho_0(y, y') \le k\sigma$ with $k \le 2\log n/\varepsilon_1$, then*

$$\left(1 - \frac{k}{2\log n}\varepsilon_1\right) t_1(y') \le t_1(y) \le \left(1 + \frac{k}{2\log n}\varepsilon_1\right) t_1(y').$$

Proof. Since $\mu \ge 128(\log n)^3\sigma/\varepsilon_1$, for any z with $\mathcal{M}_I(y,z) > r - 2\mu$, we have

$$\alpha_1(y,z) \le \left(1 - \frac{\varepsilon_1}{64(\log n)^2}\right)^{\mu/\sigma} \le \left(1 - \frac{\varepsilon_1}{64(\log n)^2}\right)^{128(\log n)^3/\varepsilon_1} \le \frac{1}{n^2}.$$

Thus,

$$|B_I(y, r-3\mu)| \le \alpha_1(y) \le |B_I(y, r-2\mu)| + \frac{1}{n}.$$

Similarly, for any z with $|\mathcal{M}_I(y,z) - r| > 3\mu$, we have

$$\beta_1(y,z) \le \left(1 - \frac{\varepsilon_1}{64(\log n)^2}\right)^{\mu/\sigma} \le \left(1 - \frac{\varepsilon_1}{64(\log n)^2}\right)^{128(\log n)^3/\varepsilon_1} \le \frac{1}{n^2},$$

and thus

$$|B_I(y, r-2\mu, r+2\mu)| \le \beta_1(y) \le |B_I(y, r-3\mu, r+3\mu)| + \frac{1}{n}.$$

If $|B_I(y, r+2\mu)|/|B_I(y, r-2\mu)| \geq 1+\varepsilon_1/2$, then

$$\frac{\beta_1(y)}{\alpha_1(y)} \geq \frac{|B_I(y, r-2\mu, r+2\mu)|}{|B_I(y, r-2\mu)|+1/n} \geq \frac{n}{n+1}\frac{|B_I(y, r-2\mu, r+2\mu)|}{|B_I(y, r-2\mu)|} \geq \frac{\varepsilon_1 n}{2(n+1)},$$

$$t_1(y) \leq \left(\frac{\varepsilon_1}{3}\frac{2}{\varepsilon_1}\left(1+\frac{1}{n}\right)\right)^{8\log n} \leq \frac{2}{n^2}.$$

If $|B_I(y, r+3\mu)/|B_I(y, r-3\mu)| \leq 1+\varepsilon_1/3$, then

$$\frac{\beta_1(y)}{\alpha_1(y)} \leq \frac{|B_I(y, r-3\mu, r+3\mu)|+1/n}{|B_I(y, r-3\mu)|} \leq \frac{|B_I(y, r-3\mu, r+3\mu)|}{|B_I(y, r-3\mu)|} + \frac{1}{n} \leq \frac{\varepsilon_1}{3} + 1/n,$$

$$t_1(y) \geq \left(\frac{\varepsilon_1}{3}\frac{1}{\varepsilon_1/3+1/n}\right)^{8\log n} \geq \left(1-\frac{3}{n\varepsilon_1}\right)^{8\log n} = 1-o(1).$$

If $\rho_0(y, y') \leq k\sigma$, then there exists a bijection $g : V_I \to V_J$ such that

1. $g(y) = y'$;

2. for any $z \in V_I$ $\mathcal{M}_J(y', g(z)) - k\sigma \leq \mathcal{M}_I(y, z) \leq \mathcal{M}_J(y', g(z)) + k\sigma$.

Then, $(1-\frac{\varepsilon_1}{64(\log n)^2})^k\alpha_1(y') \leq \alpha_1(y) \leq \frac{1}{(1-\varepsilon_1/64(\log n)^2)^k}\alpha_1(y')$ and $(1-\frac{\varepsilon_1}{64(\log n)^2})^k\beta_1(y') \leq \beta_1(y) \leq \frac{1}{(1-\varepsilon_1/64(\log n)^2)^k}\beta_1(y')$. Thus

$$\begin{aligned}\left(1-\frac{k\varepsilon_1}{2\log n}\right)t_1(y') \leq \left(1-\frac{\varepsilon_1}{64(\log n)^2}\right)^{16k\log n} t_1(y') \leq t_1(y)\\ \leq& \frac{1}{(1-\frac{\varepsilon_1}{64(\log n)^2})^{16k\log n}}t_1(y')\\ \leq& \left(1+\frac{k\varepsilon_1}{2\log n}\right)t_1(y').\end{aligned}$$

□

Definition 6.5.13. *Let $I \in \{G, H\}$. Given a vertex $y \in \mathcal{M}_I$. Let*

$$\alpha_2(y,z) = \begin{cases} 1 & \text{if } \mathcal{M}_I(y,z) \leq r-3\mu \\ (1-\frac{\varepsilon_1}{64(\log n)^2})^{(\mathcal{M}_I(y,z)-r+3\mu)/\sigma} & \text{if } \mathcal{M}_I(y,z) > r-3\mu \end{cases}$$

and $\alpha_2(y) = \sum_z \alpha_2(y,z)$. Let

$$\beta_2(y,z) = \begin{cases} 1 & \text{if } \mathcal{M}_I(y,z) \leq 1600r\log n/\varepsilon_1 + 2\mu \\ (1-\frac{\varepsilon_1}{64(\log n)^2})^{(\mathcal{M}_I(y,z)-1600r\log n/\varepsilon_1-2\mu)/\sigma} & \text{if } \mathcal{M}_I(y,z) > 1600r\log n/\varepsilon_1 + 2\mu \end{cases}$$

and $\beta_2(y) = \sum_z \beta_2(y,z)$.

Finally, let

$$t_2(y) = \begin{cases} 1 & \text{if } \beta_2(y)/\alpha_2(y) \le 3\phi \\ \left(\frac{\phi\alpha_2(y)}{\beta_2(y)}\right)^{8\log n} & \text{if } \beta_2(y)/\alpha_2(y) > 3\phi \end{cases}$$

Lemma 6.5.14. *If $\mu \ge 128(\log n)^3\sigma/\varepsilon_1$, then for any vertex $y \in V_I$*

1. *If $|B_I(y, 1600r\log n/\varepsilon_1 + 2\mu)|/|B_I(y, r-2\mu)| \ge \frac{1}{2\phi}$, then $t_2(y) \le 2/n^2$;*
2. *If $|B_I(y, 1600r\log n/\varepsilon_1 + 3\mu)|/|B_I(y, r-3\mu)| \le \frac{1}{3\phi}$, then $1 - o(1) \le t_2(y) \le 1$;*
3. *If $\rho_0(y, y') \le k\delta$ with $k \le 2\log n/\varepsilon_1$, then*

$$\left(1 - \frac{\varepsilon_1 k}{2\log n}\right) t_2(y') \le t_2(y) \le \left(1 + \frac{\varepsilon_1 k}{2\log n}\right) t_2(y').$$

Proof. Since $\mu \ge 128(\log n)^3\sigma/\varepsilon_1$, for any z with $\mathcal{M}_I(y,z) > r - 2\mu$, we have

$$\alpha_2(y,z) \le \left(1 - \frac{\varepsilon_1}{64(\log n)^2}\right)^{\mu/\sigma} \le \left(1 - \frac{\varepsilon_1}{64(\log n)^2}\right)^{128(\log n)^3/\varepsilon_1} \le \frac{1}{n^2}.$$

Thus,

$$|B_I(y, r-3\mu)| \le \alpha_2(y) \le |B_I(y, r-2\mu)| + \frac{1}{n}.$$

Similarly, for any z with $\mathcal{M}_I(y,z) > 1600r\log n/\varepsilon_1 + 3\mu$, we have

$$\beta_2(y,z) \le \left(1 - \frac{\varepsilon_1}{64(\log n)^2}\right)^{\mu/\sigma} \le \left(1 - \frac{\varepsilon_1}{64(\log n)^2}\right)^{128(\log n)^3/\varepsilon_1} \le \frac{1}{n^2},$$

and thus

$$|B_I(y, 1600r\log n/\varepsilon_1 + 2\mu)| \le \beta_2(y) \le |B_I(y, 1600r\log n/\varepsilon_1 + 3\mu)| + \frac{1}{n}.$$

If $|B_I(y, 1600r\log n/\varepsilon_1 + 2\mu)|/|B_I(y, r-2\mu)| \ge \frac{1}{2\phi}$, then

$$\frac{\beta_2(y)}{\alpha_2(y)} \ge \frac{|B_I(y, 1600r\log n/\varepsilon_1 + 2\mu)|}{|B_I(y, r-2\mu)| + 1/n} \ge \frac{n}{n+1}\frac{|B_I(y, 1600r\log n/\varepsilon_1 + 2\mu)|}{|B_I(y, r-2\mu)|} \ge \frac{n}{2\phi(n+1)},$$

$$t_2(y) \le \left(\frac{1}{3\phi} 2\phi \left(1 + \frac{1}{n}\right)\right)^{8\log n} \le \frac{2}{n^2}$$

If $|B_I(y, 1600r\log n/\varepsilon_1 + 3\mu)|/|B_I(y, r-3\mu)| \le \frac{1}{3\phi}$, then

$$\frac{\beta_2(y)}{\alpha_2(y)} \le \frac{|B_I(y, 1600r\log n/\varepsilon_1 + 3\mu)| + 1/n}{|B_I(y, r-3\mu)|} \le \frac{|B_I(y, 1600r\log n/\varepsilon_1 + 3\mu)|}{|B_I(y, r-3\mu)|} + \frac{1}{n} \le \frac{1}{3\phi} + \frac{1}{n},$$

$$t_2(y) \ge \left(\frac{1}{3\phi}\frac{3\phi n}{3\phi + n}\right)^{8\log n} \ge \left(1 - \frac{3\phi}{n + 3\phi}\right)^{8\log n} = 1 - o(1)$$

If $\rho_0(y, y') \le k\sigma$, then there exists a bijection $g : V_I \to V_J$ such that

1. $g(y) = y'$;

2. for any $z \in V_I$, $\mathcal{M}_J(y', g(z)) - k\sigma \le \mathcal{M}_I(y, z) \le \mathcal{M}_J(y', g(z)) + k\sigma$.

Then, $\left(1 - \frac{\varepsilon_1}{64(\log n)^2}\right)^k \alpha_2(y') \le \alpha_2(y) \le \frac{1}{(1-\varepsilon_1/64(\log n)^2)^k}\alpha_2(y')$ and $(1 - \frac{\varepsilon_1}{64(\log n)^2})^k \beta_2(y') \le \beta_2(y) \le \frac{1}{(1-\varepsilon_1/64(\log n)^2)^k}\beta_2(y')$ hold. Thus

$$\begin{aligned}\left(1 - \frac{\varepsilon_1 k}{2\log n}\right) t_2(y') \le \left(1 - \frac{\varepsilon_1}{64(\log n)^2}\right)^{16k\log n} t_2(y') \le t_2(y)\\ \le & \frac{1}{(1 - \frac{\varepsilon_1}{64(\log n)^2})^{16k\log n}} t_2(y')\\ \le & \left(1 + \frac{\varepsilon_1 k}{2\log n}\right) t_2(y').\end{aligned}$$

□

Definition 6.5.15. *Given a metrics $\mathcal{M}_I$ with $I \in \{G, H\}$ and a parameter r, the flow index from $y \in I$ to $z \in I$ is*

$$\mathbf{fi}_I(y, z) = \frac{\mathbf{fi}'_I(y, z)}{\sum_{w \in \mathcal{M}_I} \mathbf{fi}'_I(y, w)}$$

where

$$\mathbf{fi}'_I(y, z) = (1 - \varepsilon_1/2)^{\mathcal{M}_I(y,z)/80r} \cdot t_1(z) \cdot t_2(z).$$

Let $\mathbf{fi}_I(y) = \sum_{z \in V_I} \mathbf{fi}_I(y, z)$.

Let $\bar{\mathbf{fi}}_I(y, z) = \mathbf{fi}_I(z, y)$, and $\bar{\mathbf{fi}}_I(y) = \sum_{z \in V_I} \bar{\mathbf{fi}}_I(y, z)$.

Lemma 6.5.16. *Let $I, J \in \{G, H\}$. If two vertices $x, v \in V_I$ and $y, w \in V_J$ satisfying*

1. $\rho_1(x, y) \le k\sigma$

2. $\max\{|\mathcal{M}_I(x, v) - \mathcal{M}_J(y, w)|, |\rho_0(v) - \rho_0(w)|\} \le k\delta$

with $k \le \frac{\log n}{10\varepsilon_1}$, then $\left(1 - \frac{7k\varepsilon_1}{2\log n}\right) \mathbf{fi}_J(y, w) \le \mathbf{fi}_I(x, v) \le \left(1 + \frac{7k\varepsilon_1}{2\log n}\right) \mathbf{fi}_J(y, w)$.

Proof. Since $\rho_1(x, y) \le k\sigma$, there exists a bijection $g : V_I \to V_J$ satisfying

1. $g(x) = y$;

2. for any $z \in V_I$, $\mathcal{M}_J(y, g(z)) - k\sigma \le \mathcal{M}_I(x, z) \le \mathcal{M}_J(y, g(z)) + k\sigma$;

3. for any $z \in V_I$, $\rho_0(z, g(z)) \leq k\sigma$.

By Lemma 6.5.12 and 6.5.14, for any $z \in V_I$, $\left(1 - \frac{\varepsilon_1 k}{2\log n}\right)^2 t_1(g(z))t_2(g(z)) \leq t_1(z)t_2(z) \leq \left(1 + \frac{\varepsilon_1 k}{2\log n}\right)^2 t_1(g(z))t_2(g(z))$. Hence, for any $z \in V_I$ we have

$$\left(1 - \frac{3k\varepsilon_1}{2\log n}\right) \mathbf{fi}'_J(y, g(z)) \leq \mathbf{fi}'_I(x, z) \leq \left(1 + \frac{3k\varepsilon_1}{2\log n}\right) \mathbf{fi}'_J(y, g(z)),$$

and thus if $z \in V_I, z' \in V_J$ satisfying $\max\{|\mathcal{M}_I(z, z') - \mathcal{M}_J(z, z')|, |\rho_0(z) - \rho_0(z')|\} \leq k\delta$, then

$$\left(1 - \frac{7k\varepsilon_1}{2\log n}\right) \mathbf{fi}_J(y, z') \leq \mathbf{fi}_I(x, z) \leq \left(1 + \frac{7k\varepsilon_1}{2\log n}\right) \mathbf{fi}_J(y, z'). \tag{6.5}$$

□

Lemma 6.5.17. *Let $I, J \in \{G, H\}$. If two vertices $x \in V_I$ and $y \in V_J$ satisfying $\rho_2(x, y) \leq k\sigma$ with $k \leq \frac{\log n}{10\varepsilon_1}$, then $\left(1 - \frac{7k\varepsilon_1}{2\log n}\right) \bar{\mathbf{fi}}_J(y) \leq \bar{\mathbf{fi}}_I(x) \leq \left(1 + \frac{7k\varepsilon_1}{2\log n}\right) \bar{\mathbf{fi}}_J(y)$.*

Proof. Since $\rho_2(x, y) \leq k\delta$, there exists a bijection $g' : V_I \to V_J$ satisfying

1. $g'(x) = y$;

2. for any $z \in V_I$, $\max\{|\mathcal{M}_I(x, z) - \mathcal{M}_J(y, g'(z))|, |\rho_1(z) - \rho_1(g'(z))|\} \leq k\delta$.

By Lemma 6.5.16, for any $z \in V_I$,

$$\left(1 - \frac{7k\varepsilon_1}{2\log n}\right) \mathbf{fi}_J(g'(z), y) \leq \mathbf{fi}_I(z, x) \leq \left(1 + \frac{7k\varepsilon_1}{2\log n}\right) \mathbf{fi}_J(g'(z), y),$$

and thus $\left(1 - \frac{7k\varepsilon_1}{2\log n}\right) \bar{\mathbf{fi}}_J(y) \leq \bar{\mathbf{fi}}_I(x) \leq \left(1 + \frac{7k\varepsilon_1}{2\log n}\right) \bar{\mathbf{fi}}_J(y)$. □

Lemma 6.5.18. *For any vertex $x \in A_I$, the total internal flow from x to $V_I - S_I$ is at least $1 - O(1/n)$.*

Proof. By Lemma 6.5.12 and 6.5.14, $t_1(x) = \Omega(1)$ and $t_2(x) = \Omega(1)$. Thus, $\mathbf{fi}'_I(x, x) = \Omega(1)$. On the other hand, by Lemma 6.5.12 and 6.5.14, either $t_1(y) = O(1/n^2)$ or $t_2(y) = O(1/n^2)$ holds for any $y \in V_I - S_I$. So, $\mathbf{fi}'_I(x, y) = O(1/n^2)$. Hence $\sum_{y \in D_I} \mathbf{fi}_I(x, y) = \frac{\sum_{y \in S_I} \mathbf{fi}'_I(x,y)}{\sum_{y \in V_I} \mathbf{fi}'_I(x,y)} = O(1/n)$. □

We define the crossing flow from vertices of graph G to vertices of graph H.

Definition 6.5.19. *Given two graphs G, H with distance metrics $\mathcal{M}_G, \mathcal{M}_H$ and a parameter r. For a vertex $x \in V_G$, if there is a vertex $x' \in V_H$ matching x within distance $40r$, then the crossing flow index from x to $y \in V_H$ is*

$$\mathbf{cross}(x, y) = \mathbf{fi}(x', y).$$

If there are more than one such x', then use an arbitrary one. If there is no vertex in V_H matching x, then $\mathbf{cross}(x, y) = 0$ for any $y \in V_H$.

6.5.3 Testing label bijection

Let π be a bijection from vertices of G to vertices of H. We say π preserves crossing flow **cross** if the following two conditions holds

1. $\mathcal{M}(x, \pi(x)) \leq 1200r \log n/\varepsilon_1$
2. The distance between $\pi(x)$ and the vertex deciding the crossing flow from x to V_H is at most $1200r \log n/\varepsilon_1$ in $\mathcal{M}_J$.

Let SH_ℓ for $0 \leq \ell \leq 6 \log n/\varepsilon_1$ be set of all the vertices y in S_H with $(1+\varepsilon_1)^\ell/n^2 \leq \bar{\mathbf{fi}}_H(y) < (1+\varepsilon_1)^{\ell+1}/n^2$. We prove the following sufficient condition for the existence of bijection preserving crossing flow.

Lemma 6.5.20. *If following conditions hold,*

1. *A total weight of at most $\sqrt{n}$ vertices in V_G are at least ϕ^2-distorted for f_I with weight function $\mathbf{fi}$.*
2. *A total weight of at most $\sqrt{n}$ vertices in V_H are at least ϕ^2-distorted for f_H with weight function $\bar{\mathbf{fi}}$.*
3. *A total weight of at most $12\varepsilon_1 n$ vertices in S_G are not matched by vertices in V_H within distance $40r$ using weight function $\mathbf{fi}$.*
4. *A total weight of at most $12\varepsilon_1 n$ vertices in S_H are not matched by vertices in V_G within distance $40r$ using weight function $\bar{\mathbf{fi}}$.*
5. *For any SH_i, SH_i is not a distorted set for V_H.*

then there exists a mapping $\pi : V_G \to V_H$ such that for at least $(1-4\varepsilon)n$ vertices x in A_G,

a. $\mathcal{M}(x, \pi(x)) \leq 1200r \log n/\varepsilon_1$

b. *The distance between $\pi(x)$ and the vertex deciding the crossing flow from x to V_H is at most $1200r \log n/\varepsilon_1$ in $\mathcal{M}_J$.*

Lemma 6.5.21. *For any vertex $x \in A_I$ for $I \in \{G, H\}$, the total flow from x to $V_I - B_I(x, 320r \log n/\varepsilon_1)$ is at most $O(1/n)$.*

Proof. By Lemma 6.5.12 and 6.5.14, $t_1(x) = \Omega(1)$ and $t_2(x) = \Omega(1)$, and thus, $\mathbf{fi}'_I(x, x) = \Omega(1)$. For any $v \in V_I - B_I(x, 320r \log n/\varepsilon_1)$, we have $\mathbf{fi}'_I(x, y) \leq (1 - \varepsilon_1/2)^{320r \log n/80r\varepsilon_1} = O(1/n^2)$. Hence, the total flow from x to $V_I - B_I(x, 320r \log n/\varepsilon_1)$ is at most $O(1/n)$. □

Let A'_G be the set of vertices in A_G matched by vertices in V_J.

Lemma 6.5.22. *The total crossing flow from A'_G to S_H is at least $|A'_G| - O(1)$.*

Proof. We prove that for any vertex $x \in A'_G$, the total flow from x to $V_H - S_H$ is $O(1/n)$, and then the total flow from A'_G to $V_H - S_H$ is $O(n \cdot \frac{1}{n}) = O(1)$. Let $x' \in H$ be the vertex deciding the crossing flow from x to V_H. Since $\rho_2(x, x') \leq 2\sigma$, by Lemma 6.5.12 and 6.5.14, $t_1(x') = \Omega(1)$ and $t_2(x') = \Omega(1)$. Thus, $\mathbf{fi}'_H(x', x') = \Omega(1)$. On the other hand, by Lemma 6.5.12 and 6.5.14, either $t_1(y) = O(1/n^2)$ or $t_2(y) = O(1/n^2)$ holds for any $y \in V_H - S_H$. So, $\mathbf{fi}'_H(x', y) = O(1/n^2)$. Hence $\sum_{y \in V_H - S_H} \mathbf{cross}(x, y) = \frac{\sum_{y \in V_H - S_H} \mathbf{fi}'_H(x', y)}{\sum_{y \in V_H} \mathbf{fi}'_H(x', y)} = O(1/n)$. □

Let Ψ_H be the set of vertices $y \in S_H$ such that there is no $z' \in V_G$ matching z within distance $40r$. For every vertex $z \in S_H - \Psi_H$, let $\tau(z)$ be a vertex in V_G matching z within distance $40r$. If there is more than one possible $\tau(z)$, then use an arbitrary one. We say a pair $(x \in A'_G, z \in S_H - \Psi_H)$ is bad if

1. The distance between x and $\tau(z)$ is not distorted by f_G, and $\mathcal{M}_G(x, \tau(z)) \leq 320r \log n/\varepsilon_1$.

2. $\mathcal{M}_H(x', z) \geq \mathcal{M}_G(x, \tau(z)) + 300r$.

For a bad pair (x, z), let the expected flow from x to z be $\mathbf{fi}_G(x, \tau(z))$.

Lemma 6.5.23. *If the conditions of Lemma 6.5.20 satisfy, then the total amount of expected flow of bad pairs is at most ϕn.*

Proof. We show that if the total amount of expected flow of bad pairs is more than ϕn, then there exists a SH_ℓ such that either SH_ℓ is a distorted set of V_H, or V_G is a distorted set.

We consider a set SH_ℓ such that the total amount of expected flow for bad pairs between A'_G and $SH_\ell - \Psi_H$ is at least $\frac{\varepsilon_1 \phi n}{10 \log n}$. For any $y \in V_H$, let $\mathsf{determine}(y)$ be the set of vertices in S_G such that the flow of x is determined by y, $q_\ell(y)$ be the total amount of expected flow of bad pairs from $\mathsf{determine}(y)$ to vertices in $SH_\ell - \Psi_H$, and $\mathsf{Bad}_\ell(y)$ be the set of all the vertices $z \in SH_\ell - \Psi_H$ such that there is a $x \in \mathsf{determine}(y)$ forming bad pair with z.

For any $y \in SH_\ell - \Psi_H$, any $z \in \mathsf{Bad}_\ell(y)$, let x be a vertex in $\mathsf{determine}(y)$ such that (x, z) is a bad pair. $\mathcal{M}(y, z) \leq \mathcal{M}(y, x) + \mathcal{M}(x, \tau(z)) + \mathcal{M}(\tau(z), z) \leq 40r + \mathcal{M}_G(x, \tau(z)) + 2\delta + 40r$. For any vertex $u \in B_H(y, 80r + 8\delta)$ such that the distance between u and y is not distorted, $\mathcal{M}(u, z) \leq \mathcal{M}(u, y) + \mathcal{M}(y, z) \leq \mathcal{M}_G(x, \tau(z)) + 160r + 10\delta$. On the other hand,

$$\mathcal{M}_H(u, z) \geq \mathcal{M}_H(y, z) - \mathcal{M}_H(u, y) \geq \mathcal{M}_G(x, \tau(z)) + 300r - 80r - 8\delta = \mathcal{M}_G(x, \tau(z)) + 220r - 8\delta.$$

Hence, the distance between u and z is distorted.

Consider the case that there exists a $y \in SH_\ell - \Psi_H$ such that $q_\ell(y) \geq \sqrt{n}$. Then $|\mathsf{determine}(y)| \geq \sqrt{n}$. If there exists a vertex $x \in \mathsf{determine}(y)$ satisfying $|\{v \in \mathsf{determine}(y) : \mathcal{M}_G(x, v) \leq 80r + 4\delta\}| \geq \sqrt{n}/2$, then $|B_H(y, 80r + 8\delta)| \geq \sqrt{n}/2$ by $\rho_2(x, y) \leq 4\delta$. Since y is at most ϕ^2-distorted, at least $\sqrt{n}/4$ vertices in $B_H(y, 80r + 8\delta)$ have distance at most $80r + 10\delta$ to y in $\mathcal{M}$, and then all these vertices has distance distorted to every vertex of $\mathsf{Bad}_\ell(y)$ by f_H. On the other hand, since $q_\ell(y) \geq \sqrt{n}$ and $\bar{\mathbf{fi}}$ function is robust, $|\mathsf{Bad}_\ell(y)| \geq \max\{1, \frac{n^{5/2}}{(1+\varepsilon_1)^{\ell+2}}\}$. Using the fact that $|SH_\ell| \leq n^3/(1+\varepsilon_1)^\ell$, SH_ℓ is a distorted set of V_H. Otherwise, for every $x \in \mathsf{determine}(y)$, at least $\sqrt{n}/2 - 1$ vertices in $\mathsf{determine}(y)$ have distance distorted to x by f_G. Hence $\mathsf{determine}(y)$ is a distorted set of V_G.

Now we assume $q_\ell(y) < \sqrt{n}$ for every y. For any vertex $z \in SH_\ell - \Psi_H$, the total amount of expected flow from vertices in A_G to z is at most $\frac{(1+\varepsilon_1)^{\ell+2}}{n^2}$. Since $|SH_\ell| \leq \frac{n^3}{(1+\varepsilon_1)^\ell}$, and the total amount of expected flow to vertices in SH_ℓ is at least $\frac{\varepsilon_1 \phi n}{10 \log n}$, SH_ℓ is a distorted set of V_H. □

Definition 6.5.24. *We say the crossing flow from $x \in A_G$ to $y \in S_H - \Psi_H$ is an effective crossing flow with respect to mapping f_G and f_H if*

1. *Vertex x is at most ϕ^2-distorted, and the crossing flow from x to V_H is decided by a vertex $x' \in V_H$ matching x within distance $40r$.*

2. *Vertex y is at most ϕ^2-distorted.*

3. *The distance between x and $\tau(y)$ is not distorted by f_G.*

4. $\mathcal{M}_H(x', y) < \mathcal{M}_G(x, \tau(y)) + 300r$

5. $\mathcal{M}(x, y) \leq 400r \log n/\varepsilon_1$

For any vertex $y \in S_H$, let $\mathsf{eff}(y)$ be the total amount of effective flow from vertices in S_G to y, and $\mathsf{ub}(y) = \min\{\mathbf{fi}_H(y), \mathsf{eff}(y)\}$.

Lemma 6.5.25. *If the conditions of Lemma 6.5.20 satisfy, then $\sum_{y \in S_H} \mathsf{ub}(y) \geq (1 - 2\varepsilon)n$.*

Proof. For any vertex $y \in S_H - \Psi_H$, let $\Gamma(y)$ be the set of vertices in $B_G(\tau(y), 320r \log n/\varepsilon_1) \cap A_G$ with effective flow to y, and $\Delta(y) = (B_G(\tau(y), 320r \log n/\varepsilon_1) \cap A_G) \backslash \Gamma(y)$. We first bound $\sum_{y \in S_H - \Psi_H, x \in \Delta(y)} \mathbf{fi}_G(x, \tau(y))$.

Let P_i be the set of pairs $(x \in A_G, y \in S_H - \Psi_H)$ with $x \in \Delta(y)$ such that the crossing flow from x to y does not satisfying the i-th condition of Definition 6.5.24 for $1 \leq i \leq 5$, and $P_i' = P_i - \cup_{j<i} P_i$.

Since at most $\sqrt{n}$ vertices in V_G are at least ϕ^2-distorted, and at most $12\varepsilon_1 n$ vertices in A_G are not matched by vertices in V_H, $\sum_{(x,y) \in P_1'} \mathbf{fi}_G(x, \tau(y)) \leq \sqrt{n} + 12\varepsilon_1 n$.

Since a total weight of at most $\sqrt{n}$ vertices in V_H are at least ϕ^2-distorted using $\bar{\mathbf{fi}}$ weight function, $\sum_{(x,y) \in P_2'} \mathbf{fi}_G(x, \tau(y)) \leq \sqrt{n}$.

Since every vertex $\tau(y)$ is at most ϕ^2-distorted, $\sum_{(x,y) \in P_3'} \mathbf{fi}_G(x, \tau(y)) \leq \phi^2 n$.

If a pair $(x, y) \in P_4'$, then (x, y) is a bad pair. By Lemma 6.5.23, $\sum_{(x,y) \in P_3'} \mathbf{fi}_G(x, \tau(y)) \leq \phi n$.

For any pair (x, y) with $x \in B_G(\tau(y), 320r \log n/\varepsilon_1)$ satisfies the first four conditions of Definition 6.5.24, the fifth condition is also satisfied.

Hence,

$$\sum_{\substack{y \in S_H - \Psi_H, \\ x \in \Delta(y)}} \mathbf{fi}_G(x, y') \leq \sqrt{n} + 12\varepsilon_1 n + \sqrt{n} + \phi^2 n + \phi n \leq 13\varepsilon_1 n.$$

Let $W_H(y) = B_H(y, 320r\log n/\varepsilon_1+2\delta)\cap Z_H$ For any $y \in S_H-\Psi_H$, Since $\rho_2(y,\tau(y)) \le 4\delta$, by Lemma 6.5.16,

$$\sum_{x\in\Gamma(y)\cup\Delta(y)} \mathbf{fi}_G(x,\tau(y)) \ge (1-\varepsilon_1)\sum_{z\in W_H(y)} \mathbf{fi}_H(z,y).$$

Since $\sum_{x\in\Gamma(y)} \mathbf{cross}(x,y) \ge (1-3\varepsilon_1)\sum_{x\in\Gamma(y)} \mathbf{fi}_G(x,\tau(y))$, we have

$$\begin{aligned}
\mathsf{eff}(y) \ge& \sum_{x\in\Gamma(y)} \mathbf{cross}(x,y)\\
\ge&(1-3\varepsilon_1)\sum_{x\in\Gamma(y)} \mathbf{fi}_G(x,\tau(y))\\
\ge&(1-3\varepsilon_1)\left((1-\varepsilon_1)\sum_{z\in W_H(y)} \mathbf{fi}_H(z,y) - \sum_{x\in\Delta(y)} \mathbf{fi}_G(x,\tau(y))\right)\\
\ge&(1-4\varepsilon_1)\sum_{z\in W_H(y)} \mathbf{fi}_H(z,y) - (1-3\varepsilon_1)\sum_{x\in\Delta(y)} \mathbf{fi}_G(x,\tau(y)).
\end{aligned}$$

Hence

$$\sum_{\substack{y\in S_H-\Psi_H:\\ \mathbf{fi}_H(y)>\mathsf{eff}(y)}} \mathsf{ub}(y) \ge \sum_{\substack{y\in S_H-\Psi_H:\\ \mathbf{fi}_H(y)>\mathsf{eff}(y)}} \left((1-4\varepsilon_1)\sum_{z\in W_H(y)} \mathbf{fi}_H(z,y) - (1-4\varepsilon_1)\sum_{x\in\Delta(y)} \mathbf{fi}_G(x,\tau(y))\right).$$

If $\mathbf{fi}_H(y) \le \mathsf{eff}(y)$, $\mathsf{ub}(y) = \mathbf{fi}_H(y)$. Overall, we have

$$\sum_{y\in S_H-\Psi_H} \mathsf{ub}(y) \ge (1-4\varepsilon_1)\sum_{y\in S_H-\Psi_H, z\in W_H(y)} \mathbf{fi}_H(z,y) - (1-3\varepsilon_1)\sum_{y\in S_H-\Psi_H, x\in\Delta(y)} \mathbf{fi}_G(x,\tau(y)).$$

Since a total weight of at most $12\varepsilon_1 n$ vertices in S_H are not matched by vertices in V_G with $\bar{\mathbf{fi}}$ weight function, by Lemma 6.5.21 and Lemma 6.5.18,

$$\sum_{y\in S_H-\Psi_H, z\in W_H(y)} \mathbf{fi}_H(z,y) \ge |Z_H| - 12\varepsilon_1 n - O(1).$$

Put all together, we have We have

$$\sum_{y\in S_H-\Psi_H} \mathsf{ub}(y) \ge (1-4\varepsilon_1)[(1-\varepsilon)n - 12\varepsilon_1 n - O(1)] - (1-3\varepsilon_1)13\varepsilon_1 n \ge (1-2\varepsilon)n.$$

□

Definition 6.5.26. *We say an internal flow from $y \in V_H$ to $z \in S_H$ is an effective internal flow with respect to mapping f_H if*

1. $y \in A_H$.
2. *The distance between y and z is not distorted by f_H.*
3. $\mathcal{M}(y,z) \le 800r \log n/\varepsilon_1$

Lemma 6.5.27. *If at most $\sqrt{n}$ vertices in V_H are at least ϕ^2-distorted, then the total amount of effective internal flow is at least $(1-2\varepsilon)n$.*

Proof. Since $|V_H \backslash A_H| \le \varepsilon n$, the total flow from $V_H \backslash A_H$ to S_H is at most εn. Using the condition that at most $\sqrt{n}$ vertices in V_H are at least ϕ^2-distorted, the total flow between A_H to S_H with distance distorted is at most $\sqrt{n} + \phi^2 n$. By Lemma 6.5.21, the total flow from from A_H to S_H not satisfying the the third condition of Definition 6.5.24 is at most $O(1)$. Hence, the total effective internal flow from is at least $n - \varepsilon n - \sqrt{n} - \phi^2 n - O(1) \ge (1-2\varepsilon)n$. □

Proof of Lemma 6.5.20. We construct a function $p : A_G \times S_H \to \mathbb{R}^{\ge 0}$ such that for any $x \in A_G$, $y \in S_H$ with $p(x,y) > 0$, then $\mathcal{M}(x,y) \le 1200r \log n/\varepsilon_1$. Let $x_1, x_2, \ldots, x_{|A_G|}$ be an arbitrary order of vertices in A_G, and $y_1, y_2, \ldots, y_{|S_H|}$ be an arbitrary order of vertices in S_H. Define $p_1 : A_G \times S_H \to \mathbb{R}^{\ge 0}$ as

$$p_1(x_i,y_j) = \begin{cases} \mathbf{cross}(x_i,y_j) & \text{if the crossing flow from } x_i \text{ to } y_j \text{ is effective} \\ & \text{and } \sum_{k<i} p_1(x_k,y_j) + \mathbf{cross}(x_i,y_j) \le \mathbf{fi}_H(y_j) \\ \mathbf{fi}_H(y_j) - \sum_{k<i} p_1(x_k,y_j) & \text{if the crossing flow from } x_i \text{ to } y_j \text{ is effective} \\ & \text{and } \sum_{k<i} p_1(x_k,y_j) + \mathbf{cross}(x_i,y_j) > \mathbf{fi}_H(y_j) \\ 0 & \text{otherwise} \end{cases}$$

and $p_2 : A_G \times S_H \times S_H \to \mathbb{R}^{\ge 0}$ as

$$p_2(x_i,y_j,y_k) = \begin{cases} \mathbf{fi}_H(y_k,y_j) & \text{if the internal flow from } x_i \text{ to } y_j \text{ is effective} \\ & \text{and } \sum_{\ell<k} p_1(x_i,y_j,y_\ell) + \mathbf{fi}_H(y_k,y_j) \le \sum_x p_1(x,y_j) \\ \sum_x p_1(x,y_j) & \text{if the internal flow from } x_i \text{ to } y_j \text{ is effective} \\ -\sum_{\ell<k} p_1(x_i,y_j,y_\ell) & \text{and } \sum_{\ell<k} p_1(x_i,y_j,y_\ell) + \mathbf{fi}_H(y_k,y_j) > \sum_x p_1(x,y_j) \\ 0 & \text{otherwise} \end{cases}$$

Finally, let $p(x,y) = \sum_{z \in S_H} p_2(x,z,y)$. By the definition of effective crossing flow, effective internal flow and Lemma 6.5.25, 6.5.27, we have

1. for any $x \in A_G$, $\sum_{y \in S_H} p(x,y) \leq 1$;
2. for any $y \in S_H$, $\sum_{x \in A_G} p(x,y) \leq 1$;
3. for any $x \in A_G, y \in S_H$, if $p(x,y) > 0$, then $\mathcal{M}(x,y) \leq 1200r \log n/\varepsilon_1$;
4. $\sum_{x \in A_G, y \in S_H} p(x,y) \geq (1-4\varepsilon)n$.

Hence, function p corresponds to a fractional matching between A_G and S_H such that two vertices have non-zero weight if $\mathcal{M}(x,y) \leq 1200r \log n/\varepsilon_1$. Then the lemma follows. □

Subroutine `Testing-Label-Bijection`**:**

Input: Graph G, H; $S_{i,j,k}, T_{i,j,k} \subseteq V_G$, $C_{i,j,k}, H_{i,j,k} \subseteq V_H$ returned by Subroutine `Sparcification` with $I = G$, $J = H$ and $\mathbf{fi}$ as the weight function; $S'_{i,j,k}, T'_{i,j,k} \subseteq V_H$, $C'_{i,j,k}, H'_{i,j,k} \subseteq V_G$ returned by Subroutine `Sparcification` with $I = H$, $J = G$ and $\bar{\mathbf{fi}}$ as the weight function; parameter δ.

Output: Accept or reject.

1. Run Subroutine `Testing-Vertex-Distorted` for G with weight function $\mathbf{fi}$ with $\delta/5$.
2. Run Subroutine `Testing-Vertex-Distorted` for H with weight function $\bar{\mathbf{fi}}$ with $\delta/5$.
3. Run Subroutine `Testing-Collision` with $I = G$, $J = H$ using $\mathbf{fi}$ as weight function with $\delta/5$.
4. Run Subroutine `Testing-Collision` with $I = H$, $J = G$ using $\bar{\mathbf{fi}}$ as weight function with $\delta/5$.
5. For any SH_i, run Subroutine `Testing-Distance-Preserved-Set` with δ/n.
6. Reject if any run of the subroutines rejects, otherwise accept.

Corollary 6.5.28. *Fixing $f_G(x)$ and $f_H(x)$, if not all of the following condition hold,*

1. *A total weight of at most $\sqrt{n}$ vertices in V_G are at least ϕ^2-distorted for f_G with weight function* **fi**.
2. *A total weight of at most $\sqrt{n}$ vertices in V_H are at least ϕ^2-distorted for f_H with weight function* $\bar{\mathbf{fi}}$.
3. *A total weight of at most $12\varepsilon_1 n$ vertices in S_G are not matched by vertices in V_H within distance $40r$ using weight function* **fi**.
4. *A total weight of at most $12\varepsilon_1 n$ vertices in S_H are not matched by vertices in V_G within distance $40r$ using weight function* $\bar{\mathbf{fi}}$.
5. *For any SH_i, SH_i is not a distorted set for V_H.*

then Subroutine `Testing-Label-Bijection` *rejects with probability at least $1 - \delta/n^{3m}$.*

With Lemma 6.4.9, Lemma 6.4.11, Lemma 6.5.4, Lemma 6.5.20 and Corollary 6.5.28, Theorem 6.5.3 follows.

6.6 Sample vertices with small label distance

In this section, we present the overall algorithm for the graph isomorphism testing problem.

Before we give the overall algorithm, we show a subroutine to randomly sample pairs of vertices in two graphs with small label distance.

Subroutine `Sample-Collision-Pair`**:**

Input: Sets $T_{i,j,k}$, $S_{i,j,k}$ and $C_{i,j,k}$ with **fi** weight function in V_G.

Output: Accept or reject. If accept, also output a pair of collision (v, y).

1. Randomly sample a set T among all the sets $T_{i,j,k}$. The probability of sampling $T_{i,j,k}$ is $\frac{|S_{i,j,k}|}{|S_0|}$, where $S_0 = \cup_{i,j,k:\mathsf{weight}(T_{i,j,k}) \geq \varepsilon_2^2 n} S_{i,j,k}$.
2. Run Subroutine `Testing-Collision-`$T_{i,j,k}$ with T, and return its output.

Let $SU = \cup_{i,j,k:\mathsf{weight}(T_{i,j,k}) \geq \varepsilon_2^2 n} U_{i,j,k}$.

Lemma 6.6.1. *Subroutine* `Testing-Collision-`$T_{i,j,k}$ *rejects with probability at least* $1-\delta$ *for some set* $T_{i,j,k}$*, or Subroutine* `Sample-Collision-Pair` *satisfies following conditions*

1. *With probability at least* $1 - 17\varepsilon_1$*, Subroutine* `Sample-Collision-Pair` *returns a nice pair.*
2. *Fix a vertex* $x \in SU$*, the probability that Subroutine* `Sample-Collision-Pair` *returns a nice pair containing* x *is at most* $\frac{1+2\varepsilon}{n}$*.*
3. *For any nice pair* (x, y) *satisfying* $x \in S_{i,j,k}$*, Subroutine* `Sample-Collision-Pair` *returns* (x, y) *with probability at most* $\frac{1+2\varepsilon}{\gamma_{i,j,k} n}$
4. *For any vertex* $y \in V_H$*, Subroutine* `Sample-Collision-Pair` *returns a nice pair containing* y *with probability at most* $\frac{1000 \log^3 n}{\varepsilon_1^3 \phi n}$*.*

Proof. Since there are at most $O(\log^3 n/\varepsilon_1^3)$ $T_{i,j,k}$ sets for different i, j, k, $|S_0| \geq n - \varepsilon n - O(\log^3 n/\varepsilon_1^3)\varepsilon_2^2 n$. Hence, for any $T_{i,j,k}$, the probability of $T = T_{i,j,k}$ is at most $|S_{i,j,k}|/(1 - \varepsilon - O(\log^3 n/\varepsilon_1^3)\varepsilon_2^2)n$. By Lemma 6.4.31, the lemma holds. □

Now we consider the bijection π promised by Lemma 6.5.20. Let Y_G be the set of vertices in S_G satisfying (a) and (b) of Lemma 6.5.20.

Lemma 6.6.2. *At least one of the following conditions hold:*

1. *Subroutine* `Testing-Collision-`$T_{i,j,k}$ *rejects with probability at least* $1 - \delta$ *for some set* $T_{i,j,k}$*;*
2. *Not all the five conditions of Lemma 6.5.20 hold*
3. $|SU \cap Y_G| \geq (1 - 5\varepsilon)n$*. In addition, for any vertex* $x \in SU \cap Y_G$*, if* (x, y) *is a nice pair, then* $\mathcal{M}_H(\pi(x), y) \leq 1400r \log n/\varepsilon_1$*.*

Proof. By Lemma 6.4.25, $|SU| \geq (1-12\varepsilon_1-\varepsilon_2^2 \log^3 n/\varepsilon_1^3)n \geq (1-13\varepsilon_1)n$. By Lemma 6.5.20, $|Y_G| \geq (1 - 4\varepsilon)n$. Hence $|SU \cap Y_G| \geq (1 - 5\varepsilon)n$.

By (b) of Lemma 6.5.20, for every vertex $x \in Y_G$, there exists a vertex z which is a good/intermediate collision of $Assignto(x)$ satisfying $\mathcal{M}_H(z, \pi(x)) \leq 1200r \log n/\varepsilon_1 +$

$45r$. On the other hand, since $|W_{Assignto(x)}| = 1$, all the good/intermediate collisions of $Assignto(x)$ has distance at most $29r$ in $\mathcal{M}_H$, thus, $\mathcal{M}_H(y, \pi(x)) \leq \mathcal{M}_H(z, \pi(x)) + \mathcal{M}_H(z, y) \leq 1200r \log n/\varepsilon_1 + 45r + 29r \leq 1400r \log n/\varepsilon_1$. □

Lemma 6.6.3. *With probability $1 - 20\varepsilon$, the two pairs of collisions (x_0, y_0) and (x_1, y_1) returned by two executions of Subroutine* `Sample-Collision-Pair` *satisfy*

1. $x_0, x_1 \in SU \cap Y_G$.
2. *Both (x_0, y_0) and (x_1, y_1) are nice pairs.*
3. *$\pi(x_0)$ has same connectivity to $\pi(x_1)$ and y_1*
4. *$\pi(x_1)$ has same connectivity to $\pi(x_0)$ and y_0*
5. *y_1 has same connectivity to $\pi(x_0)$ and y_0*

Proof. For a nice pair (v, z), we define $p_{(v,z)} = \frac{1+2\varepsilon}{\gamma_{i,j,k} n}$. By Lemma 6.6.1, the probability of sampling (v, z) is at most $p_{(v,z)}$. Let N be the set of all the nice pairs (v, z) with $v \in SU \cap Y_G$. By Corollary 6.6.1 and Lemma 6.6.2, we have

$$\begin{aligned}
p(N) = & \sum_{i,j,k:\mathsf{weight}(T_{i,j,k}) \geq \varepsilon_2^2 n} \left(\sum_{\text{nice pair } (v,z) \in N : v \in SU_{i,j,k} \cap Y_G} p_{(v,z)} \right) \\
\leq & \sum_{i,j,k:\mathsf{weight}(T_{i,j,k}) \geq \varepsilon_2^2 n} |SU_{i,j,k} \cap Y_G|(1 + 2\varepsilon_1)(1 + \varepsilon_1)\gamma_{i,j,k} \cdot \frac{1 + 2\varepsilon}{\gamma_{i,j,k} n} \\
\leq & |SU \cap Y_G|(1 + 3\varepsilon)/n \\
\leq & 1 + 3\varepsilon.
\end{aligned}$$

Fix a vertex $y \in V_H$. Let N_y be the set of all the nice pairs (v, z) with $v \in SU \cap Y_G$ satisfying y has different connectivity to $\pi(v)$ and z, and denote

$$p(N_y) = \sum_{(v,z) \in N_y} p_{(v,z)}.$$

On the other hand, for any nice pair (v, z) with $v \in SU \cap Y_G$, by Lemma 6.6.2, there are at most $1400rn \log n/\varepsilon_1$ vertices in V_H have different connectivity to $\pi(v)$ and z. Hence (v, z) belongs to at most $1400rn \log n/\varepsilon_1$ different N_y for $y \in V_H$. Thus,

$$\sum_{y \in V_H} p(N_x) \leq p(N) \cdot \frac{1400rn \log n}{\varepsilon_1} \leq \frac{(1 + 3\varepsilon)1400rn \log n}{\varepsilon_1}$$

and then there are at most $\frac{(1+3\varepsilon)1400r\log n}{\varepsilon_1\varepsilon}n$ vertices y in V_H satisfying $p(N_y) > \varepsilon$. By Lemma 6.6.1 and Lemma 6.6.2, the total probability of sampling a nice pair (x_0, y_0) satisfying $x_0 \in SU \cap Y_G$ and $p(N_{\pi(x_0)}) \leq \varepsilon$ is at least

$$1 - O(\phi^{10}) - 5\varepsilon n\frac{1+2\varepsilon}{n} - \frac{1+2\varepsilon}{n}\frac{(1+3\varepsilon)1400r\log n}{\varepsilon_1\varepsilon}n \geq 1 - 6\varepsilon.$$

Now we assume (x_0, y_0) satisfies $p(N_{\pi(x_0)}) \leq \varepsilon$. Let $N_{(\pi(x_0),y_0)}$ be the set of nice pairs (x_1, y_1) satisfying at least one of following conditions:

1. $\pi(x_0)$ has distinct connectivity to $\pi(x_1)$ and y_1
2. $\pi(x_1)$ has distinct connectivity to $\pi(x_0)$ and y_0
3. y_1 has distinct connectivity to $\pi(x_0)$ and y_0

Since there are at most $1400rn\log n/\varepsilon_1$ vertices in V_H have different connectivity to $\pi(x_0)$ and y_0,

$$p(N_{(\pi(x_0),y_0)}) \leq p(N_{\pi(x_0)}) + \frac{1400rn\log n}{\varepsilon_1}(\frac{1+2\varepsilon}{n} + \frac{1000\log^3 n}{\varepsilon_1^3\phi n})) \leq 2\varepsilon.$$

Thus, with probability at least

$$1 - O(\phi^{10}) - 5\varepsilon n\frac{1+2\varepsilon}{n} - 2\varepsilon \leq 1 - 8\varepsilon$$

(x_1, y_1) is in $N_{(\pi(x_0),y_0)}$, and thus satisfies (4) and (5).

By union bound, we obtain the lemma. □

6.7 Overall algorithm

Finally, we present the main algorithm

Algorithm `Testing-Graph-Isomorphism`**:**

Input: An oracle to query edges in G and H, parameter δ

Output: Accept or reject.

1. Run Subroutine `Metric-Distance`.

2. Run Subroutine `Sparsification` to obtain sets $S_{i,j,k}, T_{i,j,k}, H_{i,j,k}, C_{i,j,k}$ with $I = G$, $J = H$ and $\mathbf{fi}$ weight function.

3. Run Subroutine `Sparsification` to obtain sets $S_{i,j,k},' T'_{i,j,k}, H'_{i,j,k}, C'_{i,j,k}$ with $I = H$, $J = G$ and $\bar{\mathbf{fi}}$ weight function.

4. Randomly sample $\lfloor 12 \log n \sqrt{n} \log(1/\delta)/\sigma \rfloor$ vertices in both G and H, denote as P_G and P_H.

5. Let $m = \lfloor 12 \log^2 n/\sigma \rfloor$. For each choice of $x_1, x_2, \ldots, x_m \in P_G$ and $y_1, y_2, \ldots, y_m \in P_H$, let $f_G(x) = e(x, x_1) \circ e(x, x_2) \circ \cdots \circ e(x, x_m)$ and $f_H(y) = e(y, y_1) \circ e(y, y_2) \circ \cdots \circ (y, y_m)$, where $e(u, v)$ for $u, v \in V_G$ is 1 iff (u, v) is an edge in G. Repeat following process with common random generator

   (a) Reject if Subroutine `Testing-Label-Bijection` for graph G and H rejects with parameter δ/n^{3m}.

   (b) Let $c = 0$.

   (c) Repeat following process $t = \lfloor \frac{\log(n^{3m}/\delta)}{\phi^5} \rfloor$ times: Run Subroutine `Sample-Collision-Pair` twice. If at least one of the execution rejects, then rejects. Otherwise, if the connectivity between x_0 and x_1 is same to the connectivity between y_0 and y_1, then increase c by 1.

   (d) If $c \geq (1 - \frac{\varepsilon_0}{2})t$, then accept.

6. Reject.

Theorem 6.7.1 (Completeness)**.** *Let G and H be two isomorphic graphs. Algorithm* `Testing-Graph-Isomorphism` *accepts with probability at least $1 - \delta$.*

Proof. Let π be an isomorphic bijection from G to H. We first show that with probability $1 - \delta$, there exists $x_1, x_2, \ldots x_m \in P_G$ and $\pi(x_1), \pi(x_2), \ldots, \pi(x_m) \in P_H$ such that for any $x, x' \in V_G$,

$$\mathcal{M}_G(x, x') - 2\sigma \leq \mathcal{M}(x, x') \leq \mathcal{M}_G(x, x') + 2\sigma,$$

and

$$\mathcal{M}_H(\pi(x),\pi(x')) - 2\sigma \leq \mathcal{M}(\pi(x),\pi(x')) \leq \mathcal{M}_H(\pi(x),\pi(x')) + 2\sigma.$$

By Lemma 6.2.1, a fraction of $1 - \frac{1}{n^{\log n}}$ out of all the sequences of vertices of length $\lfloor 12\log^2 n/\delta^2 \rfloor$ satisfying above two conditions. Let $S = \{x \in P_G : \pi(x) \in P_H\}$. If $|S| \geq \lfloor 12\log^2 n/\delta^2 \rfloor$, then with probability $1 - \frac{1}{n^{\log n}}$, there exists a sequence of vertices of length $\lfloor 12\log^2 n/\delta^2 \rfloor$ satisfying above two conditions. By Chernoff bound, with probability at least $1-\delta$, there is a sequence of vertices satisfying above two conditions.

Now we consider the execution of step 4 with respect to $x_1, x_2, \ldots x_m \in P_G$ and $\pi(x_1), \pi(x_2), \ldots, \pi(x_m) \in P_H$. By Corollary 6.5.28, step 5(a) passes with probability at least $1 - \delta/n^{3m}$.

Since π is an isomorphic mapping, for any $x_0, x_1 \in V_G$, $e(x_0, x_1)$ in G is always same to $e(\pi(x_0), \pi(x_1))$ in H. Thus, if (x_0, y_0) and (x_1, y_1) satisfying the five conditions in Lemma 6.6.3, $e(x_0, x_1)$ is same to $e(y_0, y_1)$ in H. By Lemma 6.6.3, the probability that two pairs satisfies the five conditions is at least $1 - 20\varepsilon$. Let X_i be the indicator variable that i-th execution of step 4(c) of the algorithm increase c by 1. We have $\Pr[X_i] \geq 1 - 20\varepsilon$. By Chernoff bound, with probability $1 - \delta/n^{3m}$, $c \geq (1 - \frac{\varepsilon_0}{2})t$. Thus, the algorithm accepts with probability at least $1 - O(\delta/n^{3m})$. □

Theorem 6.7.2 (Soundness)**.** *Let G and H be two graphs with distance at least ε_0. Algorithm* `Testing-Graph-Isomorphism` *rejects with probability at least $1 - \delta$.*

Proof. By Corollary 6.5.28, with probability $1 - \delta/n^{3m}$, step 5(a) rejects if not all the the five conditions are satisfied. Hence, with probability at least $1-\delta/n^{3m}$, Lemma 6.5.20 holds, and by Lemma 6.6.2, there exists a mapping π such that $|Y_G \cap SU| \geq (1 - 5\varepsilon)n$. Since G is ε_0 far from H, there are at most $(1 - \varepsilon_0)n^2$ pair of vertices $v, w \in SU \cap Y_G$ such that $e_G(v, w)$ is same to $e_H(\pi(v), \pi(w))$.

By Lemma 6.6.1 and Lemma 6.6.3, the probability of sampling two pairs $(x_0, y_0), (x_1, y_1)$ satisfying at least one of following two conditions

1. (x_0, y_0) or (x_1, y_1) are not nice pair.

2. (x_0, y_0) and (x_1, y_1) do not satisfying all of the five conditions of Lemma 6.6.3.

3. $e(x_0, x_1) = e(\pi(x_0), \pi(x_1))$.

is at most $17\varepsilon_1 + 20\varepsilon + \left(\frac{1+2\varepsilon}{n}\right)^2 (1-\varepsilon_0)n^2 \leq 1 - \varepsilon_0 + 25\varepsilon$. Hence, the probability that step 4 of the algorithm returns two pairs (x_0, y_0) and (x_1, y_1) satisfying

1. The two pairs satisfy the five conditions of Lemma 6.6.3.

2. $e(x_0, x_1) \neq e(\pi(x_0), \pi(x_1))$.

is at least $\varepsilon_0 - 25\varepsilon$. By Chernoff bound, for each run of step 4(a) to 4(d), algorithm rejects with probability at least $1 - O(\delta/n^{3m})$. By union bound the algorithm rejects with probability at least $1 - \delta$. □

Theorem 6.0.1 is obtained by combining Theorem 6.7.1 and Theorem 6.7.2.

Bibliography

[Babai and Kucera, 1979] László Babai and Ludik Kucera. Canonical labelling of graphs in linear average time. In *Foundations of Computer Science, 1979., 20th Annual Symposium on*, pages 39–46. IEEE, 1979.

[Babai and Luks, 1983] László Babai and Eugene M. Luks. Canonical labeling of graphs. In *Proceedings of the 15th Annual ACM Symposium on Theory of Computing*, pages 171–183, 1983.

[Babai and Moran, 1988] László Babai and Shlomo Moran. Arthur-Merlin games: a randomized proof system, and a hierarchy of complexity classes. *Journal of Computer and System Sciences*, 36:254–276, 1988.

[Babai and Pyber, 1994] László Babai and László Pyber. Permutation groups without exponentially many orbits on the power set. *Journal of Combinatorial Theory, Series A*, 66(1):160–168, 1994.

[Babai and Wilmes, 2013] László Babai and John Wilmes. Quasipolynomial-time canonical form for Steiner designs. In *Proceedings of the 45th ACM Symposium on Theory of Computing*, pages 261–270, 2013.

[Babai and Wilmes, 2015] László Babai and John Wilmes. Asymptotic Delsarte cliques in strongly regular graphs. 2015. in preparation.

[Babai *et al.*, 1980] László Babai, Paul Erdös, and Stanley M. Selkow. Random graph isomorphism. *SIAM Journal on Computing*, 9(3):628–635, 1980.

[Babai *et al.*, 1983] László Babai, William M. Kantor, and Eugene M. Luks. Computational complexity and the classification of finite simple groups. *24th Annual Symposium on Foundations of Computer Science, Tucson, Arizona, USA, 7-9 November 1983*, pages 162–171, 1983.

[Babai *et al.*, 2013] László Babai, Xi Chen, Xiaorui Sun, Shang-Hua Teng, and John Wilmes. Faster canonical forms for strongly regular graphs. In *Proceedings of the 54th Annual Symposium on Foundations of Computer Science*, pages 157–166, 2013.

[Babai, 1980] László Babai. On the complexity of canonical labeling of strongly regular graphs. *SIAM Journal on Computing*, 9(1):212–216, 1980.

[Babai, 1981a] László Babai. Moderately exponential bound for graph isomorphism. In *Proceedings of the International Conference on Fundamentals of Computation Theory*, pages 34–50, 1981.

[Babai, 1981b] László Babai. On the order of uniprimitive permutation groups. *Ann. of Math.*, 113(3):553–568, 1981.

[Babai, 1985] László Babai. Trading group theory for randomness. *Proceedings of the 17th Annual ACM Symposium on Theory of Computing*, pages 421–429, 1985.

[Babai, 2014] László Babai. On the automorphism groups of strongly regular graphs I. In *Proceedings of the 5th Innovations in Theoretical Computer Science*, pages 359–368, 2014.

[Babai, 2015] László Babai. Graph isomorphism in quasipolynomial time. *Manuscript*, 2015.

[Batu *et al.*, 2013] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM (JACM)*, 60(1):4, 2013.

[Boppana *et al.*, 1987] Ravi B. Boppana, Johan Håstad, and Stathis Zachos. Does co-NP have short interactive proofs? *Information Processing Letters*, 25(2):127–132, 1987.

[Cameron, 1981] Peter J. Cameron. Finite permutation groups and finite simple groups. *Bulletin of the London Mathematical Society*, 13:1–22, 1981.

[Chen *et al.*, 2013] Xi Chen, Xiaorui Sun, and Shang-Hua Teng. Multi-stage design for quasipolynomial-time isomorphism testing of Steiner 2-systems. In *Proceedings of the 45th ACM Symposium on Theory of Computing*, pages 271–280, 2013.

[Czajka and Pandurangan, 2008] Tomek Czajka and Gopal Pandurangan. Improved random graph isomorphism. *Journal of Discrete Algorithms*, 6(1):85–92, 2008.

[Fischer and Matsliah, 2008] Eldar Fischer and Arie Matsliah. Testing graph isomorphism. *SIAM J. Comput.*, 38(1):207–225, 2008.

[Goldreich *et al.*, 1991] Oded Goldreich, Silvio Micali, and Avi Wigderson. Proofs that yield nothing but their validity or all languages in NP have zero-knowledge proof system. *Journal of the ACM*, 38(1):691–729, 1991.

[Goldwasser and Sipser, 1986] Shafi Goldwasser and Michael Sipser. Private coins versus public coins in interactive proof systems. *Proceedings of the 18th Annual ACM Symposium on Theory of Computing*, pages 59–68, 1986.

[Goldwasser *et al.*, 1985] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof-systems. *Proceedings of the 17th Annual ACM Symposium on Theory of Computing*, pages 291–304, 1985.

[Higman, 1970] D. G. Higman. Coherent configurations I. *Rendiconti del Seminario Matematico della Università di Padova*, 44:1–25, 1970.

[Luks, 1982] Eugene M. Luks. Isomorphism of graphs of bounded valence can be tested in polynomial time. *Journal of Computer and System Sciences*, 25(1):42–65, 1982.

[Miller, 1978] Gary L. Miller. On the $n^{\log n}$ isomorphism technique. In *Proceedings of the 10th Annual ACM Symposium on Theory of Computing*, pages 51–58, 1978.

[Neumaier, 1979] A. Neumaier. Strongly regular graphs with smallest eigenvalue – m. *Archiv der Mathematik*, 33:392–400, 1979.

[Paninski, 2008] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[Read and Corneil, 1977] Ronald C. Read and Derek G. Corneil. The graph isomorphism disease. *J. Graph Th.*, 1(4):339–363, 1977.

[Schur, 1933] I. Schur. Zur Theorie der einfach transitiven Permutationsgruppen. *Sitzungsberichte der Preussischen Akademie der Wissenschaften*, pages 598–623, 1933.

[Spielman, 1996] Daniel A. Spielman. Faster isomorphism testing of strongly regular graphs. *Proceedings of the 28th Annual ACM Symposium on Theory of Computing,*, pages 576–584, 1996.

[Sun and Wilmes, 2015] Xiaorui Sun and John Wilmes. Faster canonical forms for primitive coherent configurations. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 693–702. ACM, 2015.

[Valiant and Valiant, 2014] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the Fifty-fifth IEEE Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2014.

[Valiant, 2011] Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.

[Weisfeiler and Leman, 1968] Boris Weisfeiler and A. A. Leman. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsiya*, 9:12–16, 1968.

[Wielandt, 1964] Helmut Wielandt. *Finite Permutation Groups*. Academic Press, Inc., 1964.

[Wilmes, 2016] John Wilmes. *Structure, automorphisms, and isomorphisms of regular combinatorial objects*. PhD thesis, 2016.

[Zemlyachenko *et al.*, 1982] Victor N. Zemlyachenko, N. M. Korneenko, and Regina I. Tyshkevich. Graph isomorphism problem. *Zapiski Nauchn. Semin. LOMI*, 118:83–158, 215, 1982.

[Zieschang, 2010] Paul-Hermann Zieschang. *Theory of association schemes*. Springer, 2010.