Electronic Colloquium on Computational Complexity, Report No. 135 (2020)

Estimation of Graph Isomorphism Distance in the Query World

Sourav Chakraborty* Arijit Ghosh* Gopinath Mishra* Sayantan Sen*

Abstract

The graph isomorphism distance between two graphs G_u and G_k is the fraction of entries in the adjacency matrix that has to be changed to make G_u isomorphic to G_k. We study the problem of estimating, up to a constant additive factor, the graph isomorphism distance between two graphs in the query model. In other words, if G_k is a known graph and G_u is an unknown graph whose adjacency matrix has to be accessed by querying the entries, what is the query complexity for testing whether the graph isomorphism distance between G_u and G_k is less than γ_1 or more than γ_2, where γ_1 and γ_2 are two constants with $0 \leq \gamma_1 < \gamma_2 \leq 1$. It is also called the tolerant property testing of graph isomorphism in the dense graph model. The non-tolerant version (where γ_1 is 0) has been studied by Fischer and Matsliah (SICOMP'08).

In this paper, we study both the upper and lower bounds of tolerant graph isomorphism testing. We prove an upper bound of $\widetilde{\mathcal{O}}(n)$ for this problem. Our upper bound algorithm crucially uses the tolerant testing of the well studied Earth Mover Distance (EMD), as the main subroutine, in a slightly different setting from what is generally studied in property testing literature.

Testing tolerant EMD between two probability distributions is equivalent to testing EMD between two multi-sets, where the multiplicity of each element is taken appropriately, and we sample elements from the unknown multi-set **with** replacement. In this paper, our (main conceptual) contribution is to introduce the problem of *(tolerant) EMD testing between multi-sets (over Hamming cube) when we get samples from the unknown multi-set **without** replacement* and to show that *this variant of tolerant testing of EMD is as hard as tolerant testing of graph isomorphism between two graphs*. Thus, while testing of equivalence between distributions is at the heart of the non-tolerant testing of graph isomorphism, we are showing that the estimation of the EMD over a Hamming cube (when we are allowed to sample **without** replacement) is at the heart of tolerant graph isomorphism. We believe that the introduction of the problem of testing EMD between multi-sets (when we get samples **without** replacement) opens an entirely new direction in the world of testing properties of distributions.

*Indian Statistical Institute, Kolkata, India

ISSN 1433-8092

1 Introduction

Graph isomorphism (GI) has been one of the most celebrated problems in computer science. Roughly speaking, the graph isomorphism problem asks whether two graphs are structure-preserving. Namely, given two graphs G_u and G_k, graph isomorphism of G_u and G_k is a bijection $\psi : V(G_u) \to V(G_k)$ such that for all pair of vertices $u, v \in V(G_u)$, the edges $\{u, v\} \in E(G_u)$ if and only if $\{\psi(u), \psi(v)\} \in E(G_k)$ [1]. One central open problem in complexity theory is whether the graph isomorphism problem can be solved in polynomial time. Recently in a breakthrough result, Babai [Bab16] proved that the graph isomorphism problem could be decided in quasi-polynomial time.

For a central problem like the graph isomorphism, naturally, one would like to understand its (and related problems) computational complexity for various models of computation. While most of the focus has been on the standard time complexity in the RAM model for various classes of graphs (and hyper-graphs), other complexity measures like space complexity, parameterized complexity, and query complexity have also been studied over the past few decades (see the Dagstuhl Report [BDST15] and PhD thesis of Sun [Sun16]).

A natural extension of the GI problem is to estimate the "graph isomorphism distance" between two graphs. In other words, given two graphs G_u and G_k, what fraction of edges are necessary to add or delete to make the graphs isomorphic.

Definition 1.1. Let $G_u = (V_u, E_u)$ and $G_k = (V_k, E_k)$ be two graphs with $|V_u| = |V_k| = n$. Given a bijection $\phi : V_u \to V_k$, the distance between the graphs G_u and G_k with respect to the bijection ϕ is

$$d_\phi(G_u, G_k) := |\{(u, v) : \text{Exactly one among } (u, v) \in E_u \text{ or } (\phi(u), \phi(v)) \in E_k \text{ holds}\}| .$$

The GRAPH ISOMORPHISM DISTANCE (or GI-distance in short) between graphs G_u and G_k is defined as $\min_{\phi: V_u \to V_k} d_\phi(G_u, G_k)/n^2$, and is denoted by $\delta_{GI}(G_u, G_k)$ (we will use $d(G_u, G_k)$ to mean $n^2 \delta_{GI}(G_u, G_k)$).

The problem of computing GI-distance between two graphs is known to be $\#P$-hard [Lin94]. The next natural question is:

What is the complexity for approximating (either by a constant additive or multiplicative factor) the graph isomorphism distance between two graphs?

In [Lin94], it was also proven that the problem of computing GI-distance between two graphs is APX-hard. So, approximating $\delta_{GI}(G_u, G_k)$ up to a constant multiplicative factor is NP-hard. In this paper, we study this problem of approximating (up to a constant additive factor) the GI-distance between two graphs in the query model.

1.1 Property Testing of Graph Isomorphism

Formally speaking, the main problem is: given two graphs G_u and G_k and an approximation parameter $\zeta \in (0, 1)$, the goal is to output an estimate α such that

$$\delta_{GI}(G_u, G_k) - \zeta \leq \alpha \leq \delta_{GI}(G_u, G_k) + \zeta.$$

[1] In a graph G, $V(G)$ and $E(G)$ denote the sets of vertices and edges in G, respectively.

In the query model, the problem is equivalent (up to a constant factor) to the tolerant property testing of graph isomorphism in the dense graph model (introduced in the work of Parnas, Ron and Rubinfeld [PRR06]). For $0 \leq \gamma < 1$, two graphs G_u and G_k, with n vertices, are called *γ-close* or *γ-far to isomorphic*[2] if $d(G_u, G_k) \leq \gamma n^2$ or $d(G_u, G_k) \geq \gamma n^2$ respectively. In *(γ_1, γ_2)-tolerant GI testing*, we are given two graphs G_u and G_k, and two parameters $0 \leq \gamma_1 < \gamma_2 \leq 1$, with the guarantee that either the graphs are γ_1-close or γ_2-far. One of the graphs (usually denoted as G_u) is accessed by querying the entries of its adjacency matrix. In contrast, the other graph (usually denoted as G_k [3]) is known to the query algorithm, and no cost for accessing the entries of the adjacency matrix of G_k is incurred. The query complexity is the number of queries (to the adjacency matrix of G_u) that are required for testing, (with correctness probability at least $2/3$ [4].), whether G_u and G_k are γ_1-close or γ_2-far. The query algorithm is assumed to have unbounded computational power.

The non-tolerant property testing version of the graph isomorphism problem (that is, when $\gamma_1 = 0$) was first studied by Fischer and Matsliah [FM08] and subsequently, Babai and Chakraborty [BC10] studied the non-tolerant property testing version of the hypergraph isomorphism problem. Like many other problems in property testing, the core difficulty in testing of GI is understanding certain properties of distributions. In the case of the non-tolerant version of GI, it was shown in [FM08] that the core problem is the testing the variation distance between two distributions. In fact, their upper bound result can be restated as: if there is a property testing algorithm, with query complexity $q(n)$ for testing equivalence between two distributions, on support size n [5], then GI can be tested using $\widetilde{\mathcal{O}}(q(n))$ queries, where the tilde hides a polylogarithmic factor of n (number of vertices). And since the query complexity for testing equivalence of distributions (from [BFF$^+$01]) is known to be $\widetilde{\mathcal{O}}(\sqrt{n})$, the query complexity for GI-testing was $\widetilde{\mathcal{O}}(\sqrt{n})$. In the lower bound proof of [FM08], there was no direct reduction of the graph isomorphism problem to the variation distance problem. But it is important to note that lower bound proofs for both of these problems use the tightness of the *birthday paradox*. So in some sense, one can say that the heart of the non-tolerant testing of GI is in the testing of variation distance between two distributions.

In this paper, we consider both the upper and lower bound on the query complexity of the tolerant version of the GI between a known and an unknown graph. Similar to the case of non-tolerant testing of GI, we show that the heart of the problem of tolerant testing of GI is in testing a certain property of distributions - but with a slight and surprising twist.

1.2 Earth Mover's Distance (EMD)

Let $H = \{0,1\}^n$ be a Hamming cube of dimension n, and p, q be two probability distributions on H. The *Earth Mover's Distance* between p and q is denoted by $EMD(p, q)$ and defined as the optimum solution to the following linear program:

$$\textbf{Minimize} \sum_{i,j \in H} f_{ij} d_H(x, y) \quad \textbf{Subject to} \sum_{j \in H} f_{ij} = p(i) \ \forall i \in H, \ \textbf{and} \sum_{i \in H} f_{ij} = q(j) \ \forall j \in H.$$

A standard way to think of sampling from any probability distribution is to consider it as a

[2]As a shorthand, rather than saying γ-close or γ-far to isomorphic, we will just say *γ-close* or *γ-far* respectively.

[3]G_u and G_k denote the unknown and known graphs respectively.

[4]The correctness probability can be made any $1 - \delta$ by incurring a multiplicative factor of $O(\log \frac{1}{\delta})$ in the query complexity

[5]Testing equivalence between two distributions means to test if the unknown distribution (from where the samples are drawn) is identical to the known distribution or is the variation distance between them more than ϵ.

multi-set of elements with appropriate multiplicities, and samples are drawn **with** replacement from that multi-set. While estimating EMD between two multi-sets, although the most natural way to access the unknown multi-set is sampling **with** replacement, we introduce the problem of tolerant EMD testing over multi-sets with the access of samples **without** replacement.

Definition 1.2 (EMD over multi-sets while sampling with and without replacement). Let S_k and S_u denote the known and the unknown multi-sets, respectively, over n-dimensional Hamming cube $H = \{0,1\}^n$ such that $|S_u| = |S_k| = n$. Consider the two distributions p_u and p_k over the Hamming cube H that are naturally defined by the sets S_u and S_k where for all $x \in H$ probability of x in p_u (and p_k) is the number of occurrences of x in S_u (and S_k) divided by n. We then define the EMD between the multi-sets S_u and S_k as

$$EMD(S_u, S_k) \triangleq n \cdot EMD(p_u, p_k).$$

The problem of estimating the EMD over multi-sets while sampling **with** (or **without**) replacement means designing an algorithm, that given any two constants β_1, β_2 such that $0 \leq \beta_1 < \beta_2 \leq 1$, and access to the unknown set S_u by sampling **with** (or **without**) replacement decides whether $EMD(S_k, S_u) \leq \beta_1 n^2$ or $EMD(S_k, S_u) \geq \beta_2 n^2$ with probability at least $2/3$.

Note that estimating the EMD over multi-sets while sampling **with** replacement is exactly same as estimating EMD between the distributions p_u and p_k with samples drawn according to p_u.

We will denote by $\text{QWR}_{\text{EMD}}(n, \beta_1, \beta_2)$ (and $\text{QWOR}_{\text{EMD}}(n, \beta_1, \beta_2)$) the number of samples **with** (or **without**) replacement required to decide the above from the unknown multi-set S_u. For ease of presentation, we will write $\text{QWOR}_{\text{EMD}}(n)$ ($\text{QWR}_{\text{EMD}}(n)$) instead of $\text{QWOR}_{\text{EMD}}(n, \beta_1, \beta_2)$ ($\text{QWR}_{\text{EMD}}(n, \beta_1, \beta_2)$) when the proximity parameters are clear from the context.

Earth Mover's Distance (EMD) is a fundamental metric over the space of distributions supported on a fixed metric space. Estimating EMD between two distributions, up to a multiplicative factor, has been extensively studied in mathematics and computer science. It is closely related to the embedding of the EMD metric into a ℓ_1 metric. Even the problem of estimation of EMD between distributions up to an additive factor has been well studied. The hardness of estimating EMD between distributions depends heavily on the structure of the domain on which the distributions are supported. In [DBNNR11], the authors have proved a lower bound of $\Omega((\Delta/\epsilon)^d)$ on the query complexity for estimating (up to an additive error of ϵ) EMD between two distributions supported on the real cube $[0, \Delta]^d$. At the same time, it is not hard to see that if the support has certain structures, estimating EMD may be easy. In this paper, we focus on the estimation of EMD between two distribution when the metric space is the Hamming cube.

As noted earlier, sample access to a probability distribution is precisely the same as uniform sampling from a multi-set **with** replacement. Thus from the results of Valiant and Valiant [VV11], it can be shown that the sample complexity for estimating the EMD between two distribution over the Hamming cube of dimension n is $\Omega(n/\log n)$. In other words, $\text{QWR}_{\text{EMD}}(n) = \Omega(n/\log n)$, and this is tight ignoring polynomial factor in $\log n$ (See Theorem B.10 of Appendix B). But what about $\text{QWOR}_{\text{EMD}}(n)$? To the best of our knowledge, the sample complexity measure when the distributions are accessed by sampling a multi-set **without** replacement has never been studied before, even for other properties of distributions. But it can be shown that: if $\text{QWOR}_{\text{EMD}}(n) = o(\sqrt{n})$, then $\text{QWR}_{\text{EMD}}(n) = o(\sqrt{n})$ (See Proposition B.7 of Appendix B). As $\text{QWR}_{\text{EMD}}(n) = \Omega(n/\log n)$, we have a lower bound of $\Omega(\sqrt{n})$ on $\text{QWOR}_{\text{EMD}}(n)$. To the best of our knowledge, there is no technique or result that would help us obtain a better lower bound than $\Omega(\sqrt{n})$ for

$\text{QWOR}_{\text{EMD}}(n)$, although a lower bound of $\Omega(n/\log n)$ exists for $\text{QWR}_{\text{EMD}}(n)$. We present the following conjecture:

Conjecture 1. *There exist two constants β_1 and β_2 with $0 < \beta_1 < \beta_2 < 1$ such that in order to decide whether $EMD(S_k, S_u) \leq \beta_1 n^2$ or $EMD(S_k, S_u) \geq \beta_2 n^2$, with probability at least $2/3$, $\Omega(\frac{n}{\log n})$ samples **without** replacement from the unknown multi-set S_u are necessary.*

One of our main contributions in this paper is introducing this complexity measure of $\text{QWOR}_{\text{EMD}}(n)$ as well as the above conjecture. In the rest of the paper, we prove the central role of this problem plays in understanding the query complexity of tolerant GI-testing.

For a formal discussion on EMD over Hamming cube, please refer to Appendix B.

1.3 Our Results

Our main result of this paper is that we prove estimating GI-distance is as hard as tolerant EMD testing over multi-sets with the access of samples **without** replacement over the unknown multi-set S_u, ignoring polynomial factors of $\log n$.

Theorem 1.3. *Let G_k and G_u denote the known and the unknown graphs on n vertices, respectively, and $Q_{GI}(G_u, G_k)$ denotes the number of adjacency queries to G_u, required by the best algorithm that takes two constants γ_1, γ_2 with $0 \leq \gamma_1 < \gamma_2 \leq 1$ and decides whether $d(G_u, G_k) \leq \gamma_1 n^2$ or $d(G_u, G_k) \geq \gamma_2 n^2$ with probability at least $2/3$. Then*

$$Q_{GI}(G_u, G_k) = \widetilde{\Theta}\left(\text{QWOR}_{\text{EMD}}(n)\right)$$

where $\widetilde{\Theta}(\cdot)$ hides polynomial factors in $\frac{1}{\gamma_2 - \gamma_1}$ and $\log n$.

Note that $\text{QWOR}_{\text{EMD}}(n) = \mathcal{O}(n)$. As a corollary to the above theorem, we obtain an upper bound on the query complexity of estimating GI-distance.

Corollary 1.4 (Upper bound of estimating GI-distance)**.** *Given a known graph G_k and an unknown graph G_u and any approximation parameter $\zeta \in (0, 1)$, there is a query algorithm that makes $\widetilde{\mathcal{O}}(n)$ queries and outputs a number α such that, with probability at least $2/3$, the following holds:*

$$\delta_{GI}(G_u, G_k) - \zeta \leq \alpha \leq \delta_{GI}(G_u, G_k) + \zeta.$$

Let us now consider the case when $\beta_1 = 0$ and $\beta_2 = \gamma$ in Definition 1.2. For this set of parameters, it can be shown that $\text{QWR}_{\text{EMD}}(n) = \Omega(\sqrt{n})$, which follows from the result on sample complexity of identity testing of distributions by Batu et.al [BFF+01]. This implies that for $\beta_1 = 0$ and $\beta_2 = \gamma$, $\text{QWOR}_{\text{EMD}}(n) = \Omega(\sqrt{n})$. Following Proposition B.7 along with Theorem 1.3, we can get an alternative proof of the following lower bound proved by Fischer and Matsliah [FM08].

Corollary 1.5 (Fischer and Matsliah [FM08])**.** *There exists a constant $\zeta \in (0, 1)$ such that any query algorithm that decides, with probability at least $2/3$, if a known graph G_k and an unknown graph G_u is isomorphic or γ-far from isomorphic, with $\gamma \leq \zeta$, must make $\Omega(\sqrt{n})$ queries.*

Although we do not have any non-trivial lower bound of tolerant EMD testing over multi-sets, we conjecture (in Conjecture 1) that the bound is tight, ignoring polynomial factors of $\log n$. Note that if Conjecture 1 is true, then following Theorem 1.3, we can say that there exists a constant $\zeta \in (0, 1)$ such that any query algorithm that estimates the GI-distance between a known graph G_k and an unknown graph G_u up to an additive factor of ζ, with correctness probability at least $2/3$, must make $\Omega(n/\log n)$ queries.

Organization of the paper. In Section 2, we discuss the proof techniques of our main results. We present the lower bound and upper bound proofs of Theorem 1.3 in Sections 3 and 4 respectively. We finally conclude in Section 5. Every theorem, lemma, and claim, whose proof has been moved to the appendix, is marked with $\star$.

Notations All graphs considered here are undirected, unweighted, and have no self-loops or parallel edges. For a graph $G(V,E)$, $V(G)$ and $E(G)$ will denote the vertex set and the edge set of G, respectively. Since we are considering undirected graphs, we write an edge $(u,v) \in E(G)$ as $\{u,v\}$. The *Hamming distance* between two points x and y in a Hamming cube $\{0,1\}^k$ will be denoted by $d_H(x,y)$.

2 Discussion on our Proofs

2.1 Discussion on the lower bound proof for query complexity

For the lower bound part of our Theorem 1.3, we give a reduction from estimating EMD of multi-sets over the Hamming cube **without** replacement to estimating the GI-distance between two graphs.

In this reduction, we have crucially used the fact that the multi-sets are composed of elements from the Hamming cube. The reduction is kind of a clever but somewhat involved gadget construction. In fact we show the lower bound for a slightly more powerful query rather than the standard adjacency matrix query that is commonly used in the dense graph model of property testing. The most interesting part of our lower bound proof is that thanks to our reduction, we get to observe the importance of the model of accessing the multi-set **without** replacement in the context of EMD testing. We are not aware of any previous work in property testing where this model of accessing a set by sampling **without** replacement has been studied.

One might compare our proof technique to the lower bound proof of (non-tolerant) testing of GI from [FM08]. In [FM08], $\Omega\left(\sqrt{n}\right)$ lower bound was proved directly (using Yao's lemma) by constructing two distributions of YES instances and NO instances - the construction of the YES and NO instances were inspired from the tightness of the birthday paradox, which was also the core idea behind the lower bound proof of the equivalence testing of two probability distributions. But, there was no direct reduction from equivalence testing of two probability distributions to GI testing. But in our lower bound proof, we establish a direct reduction to estimating EMD of multi-sets on the Hamming cube without replacement. This can be of much importance, mainly while considering other models of computation, like in the communication model. From our reduction, we can obtain an alternative proof of $\Omega(\sqrt{n})$ lower bound for the (non-tolerant) GI testing via the $\Omega(\sqrt{n})$ lower bound of the equivalence testing of distributions, as pointed out in Corollary 1.5.

2.2 Discussion on the upper bound proof for query complexity

The upper bound part of Theorem 1.3 is the main technical contribution of this paper. For the upper bound proof of the tolerant GI testing, our query algorithm is inspired by the algorithm of Fischer and Matsliah [FM08] for non-tolerant GI testing, but at the same time, our algorithm and its analysis are very different from that of [FM08] and is way more complicated and involved.

Fischer-Matsliah's algorithm is very much tuned for the non-tolerant version of GI testing. They essentially use the fact that if G_u and G_k are isomorphic, then there is a mapping between the vertex sets that makes the edge sets identical. Their algorithm tries to find whether such a mapping exists by cleverly querying induced subgraphs of G_u and checking if there exists a mapping of the queried vertices into the vertex set of G_k that can be *extended* to the whole of the vertex set. So, if for a mapping, there is an edge in G_u but not in G_k (or vice versa), it is rejected immediately. Let us first recollect Fischer-Matsliah's algorithm (FM-ALG).

FM-ALG has two phases. In the first phase, they query an induced subgraph (on $\mathcal{O}(\log^2 n)$ vertices) of G_u. The set of vertices of the induced graph is called the "*core*" set of vertices. They identify all the possible placements of the core set in the known graph G_k. Since they were only interested in the non-tolerant setting, a possible placement is a mapping of the vertices in the core set into the vertices of G_k such that the edge sets are not conflicting. The core set of vertices defines a label for each vertex in G_u: label of vertex v is the vector representing its neighbors in the core set. The collection of labels of vertices in G_u can be thought of as a distribution μ_u on the set of all possible labels. Similarly, for each placement c, of the core set into G_k, the labels of the vertices in G_k gives the distribution μ_k^c. Finally, in Phase 1, they test a "global property" of the graph by testing if for a particular placement c, the variation distance between μ_k^c and μ_u is zero or more than ϵ. This can be done simultaneously for all possible placements with query complexity $\widetilde{\mathcal{O}}(\sqrt{n})$, using (non-tolerant) testing algorithm for equivalence of distributions from [BFF$^+$01]. Only those placements that pass the test are kept as "possible placements". In Phase 2, a newly induced subgraph of G_u (on $\mathcal{O}(\log^4 n)$ vertices) is queried along with the labels of all the newly selected vertices. If there exists a possible placement that can be *extended* to a suitable placement of the new vertices, the tester outputs that G_u is isomorphic to G_k.

FM-ALG cannot be adapted to tolerant GI testing. To start with, if G_u and G_k are close, every possible mapping of the core set into vertices of G_k can be extended into a mapping of the whole vertex set such that the edge sets of G_u and G_k are close (but not necessarily identical). So, no placement can be ruled out easily. Likewise, if G_u and G_k are close, the distribution μ_u and μ_k^c may be close in variation distance (but not necessarily identical). So, one possible option is to use tolerant testing of distributions for μ_u and μ_k^c. But, the proof of correctness of the algorithm would not go through even with the tolerant testing of the equivalence of distributions. The central innovation in our upper bound result is that we use Earth Mover's Distance instead of variation distance between the distributions μ_u and μ_k^c for testing the "global property". In order to handle all these technical hurdles, our algorithm and its analysis become much more delicate and involved.

3 Lower Bound Results

In this section, we prove that it is necessary to perform $\Omega\left(\mathrm{QWoR}_{\mathrm{EMD}}(n)\right)$ many queries to the adjacency matrix of G_u to solve (γ_1, γ_2)-tolerant GI testing of G_k and G_u.

Theorem 3.1 (Restatement of the lower bound part of Theorem 1.3)**.** *Let G_k be the known and G_u be the unknown graph on n vertices, where $n \in N$ is sufficiently* large. *There exists a constant $\epsilon_{ISO} \in (0,1)$ such that for any given constants γ_1, γ_2 with $0 < \gamma_1 < \gamma_2 < \epsilon_{ISO}$, any algorithm that decides whether the graphs are γ_1-close or γ_2-far, requires $\mathrm{QWoR}_{\mathrm{EMD}}(n)$ adjacency queries to the unknown graph G_u where $\mathrm{QWoR}_{\mathrm{EMD}}$ is as defined in Definition 1.2.*

To prove Theorem 3.1, we show a reduction from tolerant GI testing to tolerant EMD testing over multi-sets when we have samples **without** replacement from the unknown multi-set.

Lemma 3.2. *Suppose there is a constant $\epsilon_0 \in \left(0, \frac{1}{2}\right)$ such that for all constants γ_1, γ_2 with $0 < \gamma_1 < \gamma_2 < \epsilon_0$ and any constant $T \in \mathbb{N}$, the following holds. There exists a (γ_1, γ_2)-tolerant tester for GI that, given a known graph G_k and an unknown graph G_u with $|V(G_u)| = |V(G_k)| = (T+1)n$, can distinguish whether $d(G_u, G_k) \leq \gamma_1 T n^2$ or $d(G_u, G_k) \geq \gamma_2 T n^2$ by performing Q adjacency queries to G_u.*

*Then, for any constants β_1 and β_2 with $0 < \beta_1 < \beta_2 < \frac{\epsilon_0}{2}$, the following holds where $\kappa = \frac{\beta_2 - \beta_1}{8}$ and $T_\kappa = \lceil \frac{18}{\kappa(2-\kappa)} \rceil$. There is a tolerant tester for EMD such that, given a known and an unknown multi-set S_k and S_u respectively, of the Hamming cube $\{0,1\}^{T_\kappa n}$ with $|S_k| = |S_u| = n$, can distinguish whether $EMD(S_k, S_u) \leq \beta_1 T_\kappa n^2$ or $EMD(S_k, S_u) \geq \beta_2 T_\kappa n^2$ with Q many samples **without** replacement from S_u.*

Remark 1. Observe that Lemma 3.2 talks about tolerant EMD testing between multi-sets with n elements over a Hamming cube of dimension $T_\kappa n$. But Theorem 3.1 states the lower bound of $\mathrm{QWoR}_{\mathrm{EMD}}(n)$, that is, of tolerant *EMD* testing of multi-sets with n elements over a Hamming cube of dimension n. However, the query complexity of *EMD* testing increases with the dimension of the Hamming cube (See Proposition B.9). So, we will be done with the proof of Theorem 3.1 by proving Lemma 3.2.

3.1 Tolerant GI to Tolerant EMD testing: Proof of Lemma 3.2

To define the necessary reduction for the proof of Lemma 3.2, we need to show the existence of a graph G_p satisfying some unique properties.

Lemma 3.3 ($\star$). *Let $\kappa \in (0,1)$ and $s \geq 3$ be given constants. Then for $C_{\kappa,s} = \lceil \frac{6s}{\kappa(2-\kappa)} \rceil$ and* sufficiently *large $n \in \mathbb{N}$* [6]*, there exists a graph G_p with $C_{\kappa,s} n$ many vertices such that the following conditions hold.*

- **(i)** *The degree of each vertex in G_p is at least $((1-\kappa)C_{\kappa,s} + 1)n - 1$.*
- **(ii)** *The cardinality of symmetric difference between the sets of neighbors of any two (distinct) vertices in G_p is at least $sn - 2$.*

The proof of Lemma 3.3 uses probabilistic method and is presented in the Appendix C.1.

Let $ALG(\gamma_1, \gamma_2, T)$ be the algorithm that takes γ_1 and γ_2 with $0 < \gamma_1 < \gamma_2 < \epsilon_0$ as input and decides whether $d(G_k, G_u) \leq \gamma_1 T n^2$ or $d(G_k, G_u) \geq \gamma_2 T n^2$, where $|V(G_k)| = |V(G_u)| = (T+1)n$. Now we show that for any two constants β_1 and β_2 with $0 < \beta_1 < \beta_2 < \frac{\epsilon_0}{2}$, $\kappa = \frac{\beta_2 - \beta_1}{8}$ and $T_\kappa = \lceil \frac{6s}{\kappa(2-\kappa)} \rceil$, there exists an algorithm $\mathcal{A}(\beta_1, \beta_2, \kappa, T_\kappa)$ that can test whether two multi-sets S_k and S_u over the $T_\kappa n$-dimensional Hamming cube have EMD less than $T_\kappa \beta_1 n^2$ or more than $T_\kappa \beta_2 n^2$ with Q many queries to the multi-set S_u. To be specific, algorithm $\mathcal{A}(\beta_1, \beta_2, \kappa, T_\kappa)$ for EMD testing will use algorithm $ALG(\gamma_1, \gamma_2, T)$ for (γ_1, γ_2)-tolerant GI such that $\gamma_1 = 2\beta_1$, $\gamma_2 = 2\beta_2 - 2\kappa$ and $T = T_\kappa$. Note that, as $0 < \beta_1 < \beta_2 < \frac{\epsilon_0}{2}$ and $\kappa = \frac{\beta_2 - \beta_1}{8}$, $0 < \gamma_1 < \gamma_2 < \epsilon_0$ holds. The details of the reduction, that is, algorithm $\mathcal{A}$ is described below.

[6]The lower bound of n is a constant that depends on κ and s.

Description of the reduction

Input: A known multi-set $S_k = \{k_1, \ldots, k_n\}$ over $H_{T_\kappa n} = \{0,1\}^{T_\kappa n}$ and query access to an unknown multi-set $S_u = \{u_1, \ldots, u_n\}$ over $H_{T_\kappa n}$.

Goal: To decide whether $EMD(S_k, S_u) \leq T_\kappa \beta_1 n^2$ or $EMD(S_k, S_u) \geq T_\kappa \beta_2 n^2$.

Construction of G_k and G_u from S_k and S_u: Let us first construct the graph G_k from S_k. G_k has $(T_\kappa + 1)n$ vertices partitioned into two parts $A_k = \{a_1, \ldots, a_n\}$ and $B_k = \{b_1, \ldots, b_{T_\kappa n}\}$. Now the edges of G_k are described as follows:

- $G_k[A_k]$ is a clique with n vertices.
- $G_k[B_k]$ is a copy of the graph $G_p(V_p, E_p)$ on $T_\kappa n$ vertices as stated in Lemma 3.3 with parameters $s = 3$, $\kappa = \frac{\beta_2 - \beta_1}{8}$ and $T_\kappa = C_{\kappa,3}$.
- For the cross edges between the vertices in A_k and B_k, we add the edge (a_i, b_j) to $E(G_k)$ if and only if the j-th coordinate of k_i is 1 for all $i \in [n]$ and $j \in [T_\kappa n]$.

Note that the graph G_k constructed above is unique for a given multi-set S_k. The graph G_u with the vertex sets $A_u = \{a'_1, \ldots, a'_n\}$ and $B_u = \{b'_1, \ldots, b'_{T_\kappa n}\}$ is constructed from the multi-set S_u in a similar fashion, but at the end, the vertices of A_u are permuted using a random permutation. So,

- $G_u[A_u]$ is a clique with n vertices.
- $G_u[B_u]$ is a copy of the graph $G_p(V_p, E_p)$ on $T_\kappa n$ vertices as stated in Lemma 3.3, with parameters $s = 3$, $\kappa = \frac{\beta_2 - \beta_1}{8}$ and $T_\kappa = C_{\kappa,3}$.
- Let us first pick a random permutation π on $[n]$. For the cross edges between the vertices in A_u and B_u, we add the edge $(a'_{\pi(i)}, b_j)$ to $E(G_u)$ if and only if the j-th coordinate of u_i is 1 for all $i \in [n]$ and $j \in [T_\kappa n]$.

Note that our final objective is to prove a lower bound on the query complexity for tolerant testing of GI, that is, when we have an adjacency query access to G_u. We will instead show that the lower bound holds even if we have the following query access, named as *A_u-neighborhood-query*: the tester can choose a vertex $a'_i \in A_u$ and in one go obtain the information about the entire neighborhood of a'_i in B_u.

Observe that the only part of G_u that is not known to the tester is the cross edges between A_u and B_u. So, in this case, the A_u-neighborhood query is way more stronger than the standard queries to G_u, and a lower bound for the A_u-neighborhood query would imply a lower bound on adjacency query.

Simulating Queries to G_u using samples drawn from S_u without replacement

Following the above discussion, we will only have to show how to simulate A_u-neighborhood queries using samples drawn from S_u **without** replacement. So, we can assume that the queries are of the form: *what are the neighbors of a'_i in B_u?* And since in each query the entire neighborhood of a'_i is obtained, the tester would pick different a'_i for every query. Note that in G_u, by construction, the vertices of A_u were permuted using a random permutation. So, from the point of view of the tester, the a'_i are just randomly drawn from A_u minus the set of a'_i already queried. In other word, the a'_i are just randomly drawn from A_u **without** replacement. Now because of the way the edges between A_u and B_u are constructed, the neighborhood of a random a'_i drawn from A_u **without** replacement is same as obtaining random samples from S_u **without** replacement.

It is also important to note that because of the randomness, the queries made by the tester are actually non-adaptive.

Description of algorithm $\mathcal{A}$ for testing $EMD(S_k, S_u)$

Run ALG on G_k and G_u with parameters $\gamma_1 = 2\beta_1$ and $\gamma_2 = 2\beta_2 - 2\kappa$. If ALG reports $d(G_k, G_u) \leq T_\kappa \gamma_1 n^2$, output that $EMD(S_k, S_u) \leq T_\kappa \beta_1 n^2$. Similarly, if ALG reports that $d(G_k, G_u) \geq T_\kappa \gamma_2 n^2$, then output $EMD(S_k, S_u) \geq T_\kappa \beta_2 n^2$.

Proof of Correctness of the reduction

To prove the correctness of the above reduction, let us first consider the following definition of SPECIAL bijection and its connection with $EMD(S_k, S_u)$.

Definition 3.4 (Special bijections)**.** A bijection ϕ from $V(G_k)$ to $V(G_u)$ is said to be SPECIAL if $\phi(A_k) = A_u$, $\phi(B_k) = B_u$ and $\phi(b_i) = b'_i$ for all $b_i \in B_k$. The set of all special bijections from $V(G_k)$ to $V(G_u)$ will be denoted by Φ, and $d_\Phi(G_k, G_u) := \min\limits_{\phi \in \Phi} d_\phi(G_k, G_u)$.

Lemma 3.5. *Let S_k, S_u be the known and unknown multi-sets, respectively. Then $d_\Phi(G_k, G_u) = 2 \cdot EMD(S_k, S_u)$.*

Proof. We will first prove that $d_\Phi(G_k, G_u) \leq 2 \cdot EMD(S_k, S_u)$.

Recall that $S_k = \{k_1, \ldots, k_n\}$ and $S_u = \{u_1, \ldots, u_n\}$ be the known and unknown multi-sets over the Hamming cube $H_{T_\kappa n} = \{0,1\}^{T_\kappa n}$. Also, note that G_u and G_k are the unknown and known graphs with vertex bipartitions A_u, B_u and A_k, B_k respectively as discussed earlier. Let $\psi : S_k \to S_u$ be an optimal bijection that realizes $EMD(S_k, S_u)$. Now, we will construct another bijection $\psi' \in \Phi$ such that $d_{\psi'}(G_k, G_u) = 2 \cdot EMD(S_k, S_u)$.

We construct the bijection $\psi' \in \Phi$ from $V(G_k)$ to $V(G_u)$ as follows: for each $i, j \in [n]$, $\psi'(a_i) = a'_j$ if and only if $\psi(k_i) = u_j$; for each $\ell \in [T_\kappa n]$, $\psi'(b_\ell) = b'_\ell$. From the construction of ψ' and by the definition of $d_{\psi'}(G_k, G_u)$ (See Definition 1.1), it is clear that $d_{\psi'}(G_k, G_u) = 2 \cdot EMD(S_k, S_u)$. Since $d_\Phi(G_k, G_u) = \min\limits_{\phi \in \Phi} d_\phi(G_k, G_u)$, we can say $d_\Phi(G_k, G_u) \leq d_{\psi'}(G_k, G_u) = 2 \cdot EMD(S_k, S_u)$.

Now we will prove the other way around, that is, we will show that $EMD(S_k, S_u) \leq \frac{d_\Phi(G_k, G_u)}{2}$ holds as well. Let $\psi \in \Phi$ be a bijection from $V(G_k) \to V(G_u)$ that realizes $d_\Phi(G_k, G_u)$. By definition of Φ, we can assume that $\psi(b_i) = b'_i$ for each $i \in [T_\kappa n]$. Now, let us consider a bijection ψ' from the multi-set S_k to S_u defined as follows: $\psi'(k_i) = u_j$ if and only if $\psi(a_i) = a'_j$ for all $i, j \in [n]$. Observe that $\sum\limits_{i \in [n]} d_H(k_i, \psi'(k_i)) = \frac{d_\psi(G_k, G_u)}{2}$. Thus, $EMD(S_k, S_u) \leq \sum\limits_{i \in [n]} d_H(k_i, \psi'(k_i)) = \frac{d_\psi(G_k, G_u)}{2} = \frac{d_\Phi(G_k, G_u)}{2}$.

Putting everything together, we have $d_\Phi(G_k, G_u) = 2 \cdot EMD(S_k, S_u)$. □

Now, using the following lemma, we will show how $d_\Phi(G_k, G_u)$ is related to $d(G_u, G_k)$, where Φ is the set of all SPECIAL bijections.

Lemma 3.6. *Let Φ be the set of all* SPECIAL *bijections from $V(G_k)$ to $V(G_u)$. Also, let $d_\Phi(G_k, G_u) = \min_{\phi \in \Phi} d_\phi(G_k, G_u)$. Then $d_\Phi(G_k, G_u) - 2\kappa T_\kappa n^2 \leq d(G_k, G_u) \leq d_\Phi(G_k, G_u)$.* [7]

Proof. Note that $d(G_k, G_u) \leq d_\Phi(G_k, G_u)$ follows from their definitions.

For the proof of the other side of the inequality, let us consider a bijection $\psi : V(G_k) \to V(G_u)$ that realizes $d(G_k, G_u)$, that is, $d(G_k, G_u) = d_\psi(G_k, G_u)$. If ψ is a bijection such that $\psi \in \Phi$, then $d_\Phi(G_k, G_u) - 2\kappa T_\kappa n^2 \leq d(G_k, G_u)$ holds. So, let us assume that $\psi \notin \Phi$. Then we will show that there exists a bijection $\phi \in \Phi$ such that $d_\phi(G_k, G_u) \leq d_\psi(G_k, G_u) + 2\kappa T_\kappa n^2$, which will imply $d_\Phi(G_k, G_u) \leq d_\psi(G_k, G_u) + 2\kappa T_\kappa n^2$, that is, $d_\Phi(G_k, G_u) - 2\kappa T_\kappa n^2 \leq d(G_k, G_u)$.

We will now present the construction of $\phi \in \Phi$ from ψ. Let us first partition the vertices of B_k, with respect to ψ, into three parts: $B_k = B_{BI} \sqcup B_{BN} \sqcup B_A$; for each $b_i \in B_{BI}$, $\psi(b_i) = b'_i$; for each $b_i \in B_{BN}$, $\psi(b_i) \in B_u$ but $\psi(b_i) \neq b'_i$; for each $b_i \in B_A$, $\psi(b_i) \in A_u$. Also, we partition the vertices of A_k into two parts: $A_k = A_A \sqcup A_B$; for each $a_i \in A_A$, $\psi(a_i) \in A_u$; for each $a_i \in A_B$, $\psi(a_i) \in B_u$. Let $|B_A| = |A_B| = x$ and $|B_{BN}| = y$, where $0 \leq x \leq n$ and $0 \leq x + y \leq T_\kappa n$. Now, we will construct the bijection $\phi \in \Phi$ (from ψ) by performing the following three steps in that order. Note that the construction of ϕ is not a part of our reduction. This is used for analysis purpose only.

Step (i) $\phi(u) = \psi(u)$ for all vertices $u \in B_{BI} \cup A_A$.

Step (ii) For each $a_i \in A_B$, $\phi(a_i) \in A_u \setminus \psi(A_A)$. Also, for each $b_i \in B_A$, $\phi(b_i) = b'_i \in B_u \setminus \psi(B_{BI})$.

Step (iii) For each $b_i \in B_{BN}$, $\phi(b_i) = b'_i$.

Observe that $\phi(A_k) = A_u$, $\phi(B_k) = B_u$ and $\phi(b_i) = b'_i$ for all $b_i \in B_k$, that is, ϕ is a SPECIAL bijection. It remains to show that

$$d_\Phi(G_k, G_u) \leq d_\psi(G_k, G_u) + 2\kappa T_\kappa n^2. \tag{1}$$

Recall that the graphs $G_k[B_k]$ and $G_u[B_u]$ are the *same* copies of $G_p(V_p, E_p)$, where $|V_p| = T_\kappa n$. Observe that

- From Lemma 3.3, the graphs $G_k[B_k]$ and $G_u[B_u]$ satisfy the following property[8]: cardinality of symmetric difference between the sets of neighbors of any two distinct vertices is at least $3n - 2$.
- Since $G_k[A_k]$ and $G_u[A_u]$ are cliques, the degree of each vertex in graphs $G_k[A_k]$ and $G_u[A_u]$ is exactly $n - 1$.

To prove $d_\Phi(G_k, G_u) \leq d_\psi(G_k, G_u) + 2\kappa T_\kappa n^2$, it will be sufficient to show that

$$d_\phi(G_k, G_u) \leq d_\psi(G_k, G_u) + 4x\,|A_k| + 4x + 2y\,|A_k| - y(3n-2). \tag{2}$$

From Equation 2, we will be done with the proof of Inequality 1 as

$$\begin{aligned} d_\psi(G_k, G_u) + 4x\,|A_k| + 4x + 2y\,|A_k| - y\,(3n-2) &= d_\psi(G_k, G_u) + 4xn + 4x - y(n-2) \\ &\leq d_\psi(G_k, G_u) + 8n^2 \leq d_\psi(G_k, G_u) + 2\kappa T_\kappa n^2. \end{aligned}$$

The last but one inequality follows from the fact that $0 \leq x \leq n$ and the last inequality follows from the fact that $T_\kappa = \lceil \frac{18}{\kappa(2-\kappa)} \rceil$. We present the proof of Inequality 2 in Appendix C.2.

The following lemma completes the proof of Lemma 3.2.

[7] Note that this relation does not hold in general. However this is true for the graphs G_k and G_u constructed in the reduction.

[8] Note that we are using Lemma 3.3 with parameters $s = 3$, $\kappa = \frac{\beta_2 - \beta_1}{8}$ and $T_\kappa = C_{\kappa,3}$.

Lemma 3.7. *The described algorithm $\mathcal{A}$ for EMD, that uses Algorithm* ALG *on G_k and G_u with parameters γ_1 and γ_2 as a subroutine, determines whether $EMD(S_k, S_u) \leq \beta_1 T_\kappa n^2$ or $EMD(S_k, S_u) \geq \beta_2 T_\kappa n^2$ with probability at least $2/3$, where $\gamma_1 = 2\beta_1$, $\gamma_2 = 2\beta_2 - 2\kappa$.*

Proof. By the assumption of the existence of algorithm ALG that decides whether $d(G_k, G_u) \leq T_\kappa \gamma_1 n^2$ or $d(G_k, G_u) \geq T_\kappa \gamma_2 n^2$, we will be done with the proof by showing the followings.

(i) If $EMD(S_k, S_u) \leq T_\kappa \beta_1 n^2$, then $d(G_k, G_u) \leq T_\kappa \gamma_1 n^2$,

(ii) If $EMD(S_k, S_u) \geq T_\kappa \beta_2 n^2$, then $d(G_k, G_u) \geq T_\kappa \gamma_2 n^2$.

We will first prove (i). From Lemma 3.5, we have $d_\Phi(G_k, G_u) = 2 \cdot EMD(S_k, S_u)$, where Φ is the set of all SPECIAL bijections from $V(G_k)$ to $V(G_u)$. So, $EMD(S_k, S_u) \leq T_\kappa \beta_1 n^2$ implies $d_\Phi(G_k, G_u) \leq 2T_\kappa \beta_1 n^2 = T_\kappa \gamma_1 n^2$. Now, following the definition of SPECIAL bijections (Definition 3.4) and Lemma 3.6, we can say that $d(G_k, G_u) \leq d_\Phi(G_k, G_u) \leq T_\kappa \gamma_1 n^2$.

Now, for the proof of (ii), considering the fact that $d_\Phi(G_k, G_u) = 2 \cdot EMD(S_k, S_u)$ as above, we can say that $EMD(S_k, S_u) \geq T_\kappa \beta_2 n^2$ implies $d_\Phi(G_k, G_u) \geq 2T_\kappa \beta_2 n^2$. From Lemma 3.6, it follows that $d_\Phi(G_k, G_u) - 2\kappa T_\kappa n^2 \leq d(G_k, G_u)$. Thus, $d(G_k, G_u) \geq T_\kappa(2\beta_2 - 2\kappa)n^2 = T_\kappa \gamma_2 n^2$. □

□

4 Query Algorithm for Tolerant Graph Isomorphism Testing

In this section, we prove the following theorem.

Theorem 4.1. *(Restatement of the upper bound part of Theorem 1.3) Let G_k and G_u be the known and unknown graphs, respectively. There exists an algorithm that takes parameters γ_1 and γ_2 as input such that $0 \leq \gamma_1 < \gamma_2 \leq 1$, performs $\widetilde{\mathcal{O}}\left(\mathrm{QWoR}_{\mathrm{EMD}}(n)\right)$ many queries to the adjacency matrix of G_u for appropriate β_1 and β_2 depending on γ_1 and γ_2, and decides whether $d(G_u, G_k) \leq \gamma_1 n^2$ or $d(G_u, G_k) \geq \gamma_2 n^2$, with probability at least $2/3$. Here $\widetilde{\mathcal{O}}(\cdot)$ hides a polynomial factor in $\frac{1}{\beta_2 - \beta_1}$ and $\log n$.*

Remark 2. The theorem stated above works for any γ_1, γ_2 such that $0 \leq \gamma_1 < \gamma_2 \leq 1$. However, for simplicity of representation, we have assumed $\gamma_2 \geq 11\gamma_1$.

Remark 3. Note that Theorem 4.1 can also be stated in terms of $\mathrm{QWR}_{\mathrm{EMD}}(n)$ as $\mathrm{QWoR}_{\mathrm{EMD}}(n) \leq \mathrm{QWR}_{\mathrm{EMD}}(n)$ as we can simulate samples **with** replacement when we have query access to samples **without** replacement (See Proposition B.5).

Our algorithm for tolerant GI testing, as stated in Theorem 4.1, uses a special kind of tolerant *EMD* tester over multi-sets: we know t many multi-sets, one multi-set is unknown and two parameters ϵ_1 and ϵ_2 are given; the objective is to test tolerant *EMD* of each known multi-set with the unknown one. The following theorem gives us the special *EMD* tester.

Theorem 4.2. *Let $H = \{0,1\}^n$ be a n-dimensional Hamming cube. Let $\{S_k^i : i \in [t]\} \cup \{S_u\}$ denote the multi-sets with n elements from H where $\{S_k^i : i \in [t]\}$ denote the set of t many known multi-sets and S_u denotes the unknown multi-set. There exists an algorithm* ALG-EMD *that takes two proximity parameters ϵ_1, ϵ_2 with $0 \leq \epsilon_1 < \epsilon_2 \leq 1$ and a $\delta \in (0,1)$ as input and decides whether $EMD(S_u, S_k^i) \leq \epsilon_1 n^2$ or $EMD(S_u, S_k^i) \geq \epsilon_2 n^2$, with probability at least $1 - \delta$, for each $i \in [t]$. Moreover,* ALG-EMD *uses $\mathrm{QWoR}_{\mathrm{EMD}}(n) \cdot \mathcal{O}\left(\log \frac{t}{\delta}\right)$ many samples **without** replacement from S_u.*

The above theorem follows from the definition of $\text{QWoR}_{\text{EMD}}(n)$ (See Definition 1.2) along with union bound and standard argument for amplifying the success probability.

Remark 4. The algorithm of Theorem 4.1, to be discussed in Section 4.1, formulates a tolerant *EMD* instance of multi-sets having n elements in $H = \{0,1\}^d$, where $d = \mathcal{O}(\log n/(\gamma_2 - \gamma_1))$. But ALG-EMD is an algorithm for tolerant *EMD* testing between two multi-sets having n elements in $\{0,1\}^n$. This is not a problem as the query complexity of *EMD* is an increasing function in dimension (See Proposition B.9 in Appendix B). Moreover, the algorithm in Section 4.1 calls ALG-EMD with parameters $\epsilon_1 = (\gamma_1 + \frac{\gamma_2-\gamma_1}{2000})$, $\epsilon_2 = \gamma_2/5$, $t = 2^{\mathcal{O}(\log^2 n/(\gamma_2-\gamma_1))}$ and δ is a suitable constant depending upon γ_1 and γ_2, where γ_1 and γ_2 are parameters as stated in Theorem 4.1. So, each call to ALG-EMD, in our context, makes $\widetilde{\mathcal{O}}\left(\text{QWoR}_{\text{EMD}}(n)\right)$ many queries.

4.1 Query Algorithm for Tolerant Graph Isomorphism

For our algorithm, we need the following definitions of *label* and *embedding*.

Definition 4.3. (*Label* of a vertex) Given a graph G and $C \subset V(G) = \{c_1, \dots c_{|C|}\}$, the C-labelling of $V(G)$ is a function $\mathcal{L}_C : V(G) \to \{0,1\}^{|C|}$ such that the i-th entry of $\mathcal{L}_C(v)$ is 1 if and only if v is a neighbor of $c_i \in C$. Also, $\mathcal{L}_C(v)$ is referred as the label of v under C-labelling of $V(G)$.

Definition 4.4. (*Embedding* of a Vertex Set into another Vertex Set) Let G_u and G_k be two graphs. Consider $A \subseteq V(G_u)$ and $B \subseteq V(G_k)$ such that $|A| \leq |B|$. An injective mapping η from A to B is referred as an *embedding* of A into B.

Now we present our query algorithm **TolerantGI(**G_u**,** G_k**,** γ_1**,** γ_2**)** that comprises three phases. Before proceeding to the formal description, we first give technical overview to get a flow of our algorithm.

Technical Overview

In Phase 1, we first choose a $\mathcal{O}(\frac{1}{\gamma_2-\gamma_1})$ size collection of random subset of vertices, i.e, *coresets* $\mathcal{C}_u$ from the unknown graph G_u where each $C_u \in \mathcal{C}_u$ is of size $\mathcal{O}(\frac{\log n}{\gamma_2-\gamma_1})$. Thereafter we find all embeddings of C_u inside the known graph G_k. Let the embeddings be $\eta_1, \eta_2, \dots, \eta_J$ where $C_k^i = \eta_i(C_u)$. Now each C_u (as well as each C_k^i) defines a label distribution of the vertices of G_u (as well as G_k). Let us denote the set of labels as X_{C_u} (and $Y_{C_k^i}$). Now we test if the *EMD* between X_{C_u} and $Y_{C_k^i}$ is close or far for each $i \in [J]$. We keep only those (C_u, η_i) for Phase 2 such that $EMD(X_{C_u}, Y_{C_k^i}) \leq (\gamma_1 + \frac{\gamma_2-\gamma_1}{2000}) n\, |C_u|$.

Although Phase 1 of our algorithm is similar to the algorithm of [FM08], there is a striking difference. Since the authors of [FM08] were testing the non-tolerant version of graph isomorphism, they were testing the identity of the label distributions of X_{C_u} and $Y_{C_k^i}$. However, since we are solving the tolerant version of the problem, we need to allow some error among the label distributions. We need to pass only those placements of C_u that under *good bijections* do not produce much error and testing of tolerant *EMD* fits exactly for this purpose.

In Phase 2, we choose $\mathcal{O}\left(\frac{\log^2 n}{(\gamma_2-\gamma_1)^3}\right)$ many vertices from the unknown graph G_u randomly and call it W. We further find the labels of all the vertices of W under C_u-labelling by querying the corresponding entries of G_u for each C_u that has passed Phase 1. Then, we try to match the vertices of W to the set of possible labels $\{l_1, l_2, \ldots, l_t\}$ of the vertices of G_k under C_k^i-labelling where $C_k^i = \eta_i(C_u)$, for those η_i that have passed Phase 1. Ideally, we would like to find a mapping $\psi : W \to \{l_1, l_2, \ldots, l_t\}$ such that the total distance between the labels of the matched vertices is not too large. If no such ψ is possible, we reject the current embedding and try some other embedding that has passed Phase 1.

In Phase 3, we construct a random partial bijection $\hat{\phi} : W \to V(G_k)$ that maps the vertices of W to the vertices of G_k while preserving the labels according to ψ. We achieve this by mapping each $w \in W$ to one vertex of G_k randomly that has same label as determined by ψ. Finally, we randomly pair the vertices of W and find the fraction of edge mismatches between the paired up vertices of W and $\hat{\phi}(W)$. If this fraction is less than $5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)$, we accept and say that G_u and G_k are γ_1-close. If there is no such embedding of any $C_u \in \mathcal{C}_u$ that achieves this, we report that G_u and G_k are γ_2-far.

Formal Description of TolerantGI(G_u, G_k, γ_1, γ_2):

The three phases of our algorithm are as follows:

4.1.1 Phase 1

The first phase of our algorithm consists of the following three steps.

Step 1 First we sample a collection $\mathcal{C}_u$ of $\mathcal{O}\left(\frac{\log n}{\gamma_2 - \gamma_1}\right)$ sized random subsets of $V(G_u)$ with $|\mathcal{C}_u| = \mathcal{O}(\frac{1}{\gamma_2 - \gamma_1})$. We perform **Step 2** and **Step 3** for each $C_u \in \mathcal{C}_u$.

Step 2 We determine all possible embeddings, that is, $\eta_1, \ldots, \eta_J$, of C_u into $V(G_k)$, where $J = \binom{n}{\mathcal{O}(\log n/(\gamma_2 - \gamma_1))} \leq 2^{\mathcal{O}(\log^2 n/(\gamma_2 - \gamma_1))}$. For each $i \in [J]$, let C_k^i be the set of images of C_u under the i-th embedding of C_u into $V(G_k)$, that is, $C_k^i = \eta_i(C_u)$. For all $i \in [J]$, we construct the multi-set $Y_{C_k^i}$ that contains C_k^i-labellings of all the vertices of G_k.

Step 3 Now for each vertex $v \in V(G_u)$, there is a C_u-labelling of v. Let X_{C_u} be the multi-set of C_u-labellings of all the vertices in $V(G_u)$. However, X_{C_u} is unknown to the algorithm. We call ALG-EMD (as stated in Theorem 4.2) by setting parameters as described in Remark 4 to decide whether $EMD(X_{C_u}, Y_{C_k^i}) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000})n\,|C_u|$ or $EMD(X_{C_u}, Y_{C_k^i}) \geq \gamma_2 n\,|C_u|\,/5$, for each $i \in [J]$. Let us pair up C_u's and their accepted embeddings into G_k and call the set Γ, that is,

$$\Gamma = \left\{ (C_u, \eta_i) \mid \text{ALG-EMD } \textit{decides } EMD(X_{C_u}, Y_{C_k^i}) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000})n\,|C_u| \right\}.$$

Note that, at the end of the **Phase 1**, we have Γ with $|\Gamma| \leq |\mathcal{C}_u| \cdot 2^{\mathcal{O}(\log^2 n/(\gamma_2 - \gamma_1)} = \mathcal{O}\left(2^{(\log^2 n/(\gamma_2 - \gamma_1))}\right)$. By the description of **Step 3** above, **Phase 1** of our algorithm calls ALG-EMD $\mathcal{O}(|\mathcal{C}_u|)$ times, once for each $C_u \in \mathcal{C}_u$. So, setting $\delta = \frac{1}{9|\Gamma|}$ in Theorem 4.2, we obtain the following observation about Γ that will be used to prove the soundness of our algorithm.

Observation 4.5. Consider Γ, the set of accepted embeddings that have passed **Phase 1** paired with corresponding C_u, as defined above. Then

$$\mathbb{P}\left(\forall\ (C_u, \eta_i) \in \Gamma, EMD(X_{C_u}, Y_{C_k^i}) \leq \gamma_2 n\,|C_u|\,/5\right) \geq \frac{8}{9}.$$

4.1.2 Phase 2

In the second phase, the algorithm performs the following two steps.

Step 1 We sample a subset W of $\mathcal{O}(\log^2 n/(\gamma_2-\gamma_1)^3)$ vertices randomly from G_u.

Step 2 For each $(C_u, \eta_i) \in \Gamma$ that has passed **Phase 1**, we perform the following steps:

- **(i)** We find the $C_k^i = \eta_i(C_u)$-labelling of the vertices of G_k. Let $l_1, \ldots, l_t$ be the labels of the vertices where $t = 2^{|C_k^i|}$ and $V_j \subseteq V(G_k)$ be the set of vertices with label l_j.
- **(ii)** We define a matrix M of size $|W| \times 2^{|C_k^i|}$ where each row represents the label of a vertex $w \in W$ and each column represents one of the possible C_k^i-labelling of $V(G_k)$ [9]. The (i,j)-th entry of M is defined as: $M_{ij} = d_H(\mathcal{L}_{C_u}(w_i), l_j)$.
- **(iii)** We choose a function $\psi : W \to \{l_1, \ldots l_t\}$ randomly satisfying

$$\sum_{w \in W} d_H(\mathcal{L}_{C_u}(w), \psi(w)) \leq \frac{2\gamma_2}{5}\,|C_u|\,|W| \;\text{ and }\; \left|\{w : \psi(w) = l_j\}\right| \leq \left|V_j\right| \forall\, j \in [t]. \tag{3}$$

Let Γ_W be the set of tuples such that

$$\Gamma_W = \{(C_u, \eta_i, \psi) : (C_u, \eta_i) \in \Gamma \text{ and } \psi \text{ satisfies Equation (3)}\}.$$

Like Observation 4.5, the following observation about the set Γ_W will be used to prove the soundness of our algorithm.

Observation 4.6. $|\Gamma_W| \leq |\Gamma| \leq 2^{\mathcal{O}(\log^2 n/(\gamma_2-\gamma_1))}$. Moreover, any (C_u, η_i, ψ) that has passed this phase satisfies Equation (3).

4.1.3 Phase 3

The third phase of our algorithm comprises the following four steps.

Step 1 We randomly pair up the vertices of W. Let $\{(a_1, b_1), \ldots, (a_p, b_p)\}$ be the pairs of the vertices, where $p = \mathcal{O}(\log^2 n/(\gamma_2-\gamma_1)^3)$. We now determine which (a_i, b_i) pairs form edges in G_u by querying the corresponding entries of the adjacency matrix of G_u.

Step 2 For each $(C_u, \eta_i, \psi) \in \Gamma_W$ that has passed **Phase 2**, we perform **Step 3** and **Step 4** as follows.

Step 3 We choose an embedding $\hat{\phi} : W \to V(G_k)$ randomly, satisfying $\hat{\phi}(w) \in V_j$ if and only if $\psi(w) = l_j$ and modulo permutation of the vertices in V_j for all $j \in [t]$. In other words, we map each $w \in W$ to a vertex in G_k randomly having $\psi(w) = l_j$ as its C_k^i-labelling in G_k.

[9] Let $C_u = \left\{x_1, \ldots, x_{\mathcal{O}(\log n/(\gamma_2-\gamma_1))}\right\}$. Note that for each $w_i \in W$, $\mathcal{L}_{C_u}(w_i) \in \{0,1\}^{\mathcal{O}(\log n/(\gamma_2-\gamma_1))}$ such that the j-th coordinate is 1 if and only if w_i is a neighbour of x_j, where $i \in \left[\mathcal{O}(\log^2 n/(\gamma_2-\gamma_1)^3)\right]$ and $j \in [\mathcal{O}(\log n/(\gamma_2-\gamma_1))]$. Similarly, $l_j \in \{0,1\}^{\mathcal{O}(\log n/(\gamma_2-\gamma_1))}$ such that the i-th coordinate of l_j is 1 if and only if $\eta(x_i)$ is a neighbour of $v \in V_j$, where $j \in \left[2^{|C_k^i|}\right]$.

Step 4 We find the fraction $\zeta(C_u, \eta_i, \psi, \hat{\phi}) = \left|\{(a_i, b_i) : \mathbb{1}_{(a_i,b_i)} = 1\}\right| / p$, where $\mathbb{1}_{(a_i,b_i)} = 1$ if exactly one among $(a_i, b_i) \in E(G_u)$ and $(\hat{\phi}(a_i), \hat{\phi}(b_i)) \in E(G_k)$ holds. If $\zeta(C_u, \eta_i, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)$, then **HALT and REPORT** that G_u and G_k are γ_1-close.

While executing **Step 3** and **Step 4** for each tuple in Γ_W, if we did not **HALT**, then we **HALT** now and **REPORT** that G_u and G_k are γ_2-far.

Observation 4.7. (i) The number of times our algorithm executes **Step 2**, **Step 3** and **Step 4** is at most $|\Gamma_W| \leq 2^{\mathcal{O}(\log^2 n/(\gamma_2 - \gamma_1))}$.

(ii) If there exists a (C_u, η_i, ψ) such that $\zeta(C_u, \eta_i, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)$, then our algorithm reports that G_u and G_k are γ_1-close. Otherwise, G_u and G_k are reported to be γ_2-far.

4.2 Proof of Correctness

To prove the correctness of our algorithm, we need to show the following three properties:

Completeness Property If G_u and G_k are γ_1-close to isomorphic, then our algorithm reports the same with probability at least 2/3.

Soundness Property If G_u and G_k are γ_2-far from isomorphic, then the algorithm reports the same with probability at least 2/3.

Query Complexity The query complexity of our algorithm is $\widetilde{\mathcal{O}}(n)$.

4.2.1 Proof of Completeness Property

In order to prove the completeness property as described above, we will first prove some claims. Finally, combining the claims, we would conclude the completeness property of our algorithm.

We will first prove that there exists a $C_u \in \mathcal{C}_u$ considered in **Step 1** of **Phase 1** of the algorithm and a corresponding embedding $\eta_i : C_u \to V(G_k)$ in **Step 2** of **Phase 1** such that $\text{EMD}\left(X_{C_u}, Y_{C_k^i}\right) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000})n\,|C_u|$ holds with probability at least 20/21, where $C_k^i = \eta_i(C_u)$.

Claim 4.8. *Let $\phi : V(G_u) \to V(G_k)$ be a bijection such that $d_\phi(G_u, G_k) \leq \gamma_1 n^2$. Then there exists a $C_u \in \mathcal{C}_u$ and an embedding $\eta_i : C_u \to V(G_k)$ such that the following hold with probability at least 20/21.*

- *$\forall v \in C_u$, we have $\eta_i(v) = \phi(v)$, and*
- $\text{EMD}\left(X_{C_u}, Y_{C_k^i}\right) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000})n\,|C_u|$

Note that $C_k^i = \eta_i(C_u)$ and $Y_{C_k^i}$ is set of C_k^i-labelling of $V(G_k)$.[10]

Proof. Consider a particular $C_u \in \mathcal{C}_u$ and an embedding $\eta_i : C_u \to V(G_k)$ such that $\eta_i(v) = \phi(v)$ for all $v \in C_u$. Note that this embedding η_i is considered in **Step 2** of **Phase 1** of the algorithm. Now we will show that $\text{EMD}\left(X_{C_u}, Y_{C_k^i}\right) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000})n\,|C_u|$ holds with probability at least a constant, to be specified later, that depends upon γ_1 and γ_2, where $C_k^i = \eta_i(C_u)$.

[10] C_k^i and $Y_{C_k^i}$ are defined in **Step 2** of **Phase 1**.

We know that $d_\phi(G_u, G_k) \leq \gamma_1 n^2$ and by Definition A.2, we have

$$\sum_{x \in V(G_u)} \left|\textsc{Decider}_\phi(x)\right| \leq \gamma_1 n^2.$$

Thus,

$$\mathbb{E}\left[\sum_{x \in V(G_u)} \left|\textsc{Decider}_\phi(x) \cap C_u\right|\right] \leq \gamma_1 n \left|C_u\right|. \tag{4}$$

From Definition A.2, we can say that

$$\begin{aligned} \text{EMD}\left(X_{C_u}, Y_{C_k^i}\right) &= \min_{f: V(G_u) \to V(G_k)} \sum_{x \in V(G_u)} \left|\textsc{Decider}_f(x) \cap C_u\right| \\ &\leq \sum_{x \in V(G_u)} \left|\textsc{Decider}_\phi(x) \cap C_u\right| \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[\text{EMD}\left(X_{C_u}, Y_{C_k^i}\right)\right] &\leq \mathbb{E}\left[\sum_{x \in V(G_u)} \left|\textsc{Decider}_\phi(x) \cap C_u\right|\right] \\ &\leq \gamma_1 n \left|C_u\right| \quad \text{(From Equation 4)} \end{aligned}$$

Using Markov inequality, we can say that

$$\mathbb{P}\left(\text{EMD}\left(X_{C_u}, Y_{C_k^i}\right) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000}) n \left|C_u\right|\right) \geq 1 - \frac{\gamma_1}{\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000}}.$$

Note that $|\mathcal{C}_u| = \mathcal{O}(\frac{1}{\gamma_2 - \gamma_1})$ and we have been arguing for a particular $C_u \in \mathcal{C}_u$. So, taking $|\mathcal{C}_u|$ suitably, we get a C_u and an embedding $\eta_i : C_u \to V(G_k)$ satisfying the properties mentioned in the statement of this claim with probability at least $20/21$. □

The above claim discusses about the existence of a $C_u \in \mathcal{C}_u$ and its embeddings satisfying above mentioned desired properties. Now we discuss how our algorithm determines all $C_u \in \mathcal{C}_u$ that satisfy the properties. Note that **Step 3** of **Phase 1** of our algorithm calls ALG-EMD. Following the correctness of ALG-EMD (Theorem 4.2), we determine all embeddings $\eta_i : C_u \to V(G_k)$ such that $\text{EMD}\left(X_{C_u}, Y_{C_k^i}\right) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000}) n \left|C_u\right|$ holds with probability at least $20/21$. The discussion in this paragraph is formalized in the following claim.

Claim 4.9. *Let $C_u \in \mathcal{C}_u$ and $\eta_1, \ldots, \eta_J$ be the all possible embeddings of C_u into $V(G_k)$. Then* **Step 3** *of* **Phase 1** *can determine the set $\Gamma = \{(C_u, \eta_i) \mid EMD(X_{C_u}, Y_{C_k^i}) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000}) n \left|C_u\right|\}$ with probability at least $20/21$. Note that $C_k^i = \eta_i(C_u)$, X_{C_u} is the set of C_u-labelling of $V(G_u)$ and $Y_{C_k^i}$ is set of C_k^i-labelling of $V(G_k)$.*

As we are considering the case that G_u and G_k are γ_1-close to being isomorphic, from Claim 4.8, we can assume that there is an appropriate $(C_u, \eta_i) \in \Gamma$ such that $\text{EMD}\left(X_{C_u}, Y_{C_k^i}\right) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000}) n \left|C_u\right|$. Now we will prove that there exists a function $\psi : W \to \{l_1, \ldots, l_t\}$ as considered in **Step 2 (iii)** in **Phase 2** of our algorithm such that Equation (3) holds with probability at least $20/21$.

Claim 4.10. *Let us assume that $\phi : V(G_u) \to V(G_k)$ be a bijection such that $d_\phi(G_u, G_k) \leq \gamma_1 n^2$ and $(C_u, \eta_i) \in \Gamma$ where $C_u \in \mathcal{C}_u$ and $\eta_i : C_u \to V(G_k)$ be an embedding such that*

- *$\forall v \in C_u$ we have $\eta_i(v) = \phi(v)$, and*
- *$\mathrm{EMD}\left(X_{C_u}, Y_{C_k^i}\right) \leq (\gamma_1 + \frac{\gamma_2-\gamma_1}{2000})n\,|C_u|$ where $C_k^i = \eta_i(C_u)$.*

Also, let $\{\ell_1, \ldots, \ell_t\}$ be the all possible C_k^i-labellings of $V(G_k)$, where $t = \left\lceil 2^{|C_k^i|}\right\rceil$. Then there exists a mapping $\psi : W \to \{l_1, \ldots, l_t\}$ such that the following hold with probability at least $20/21$.

(i) $\sum_{w\in W} d_H(\mathcal{L}_{C_u}(w), \psi(w)) \leq \frac{2\gamma_2}{5}\,|C_u|\,|W|$, *and*

(ii) $\forall\, j \in [t]$, *we have* $\left|\{w : \psi(w) = l_j\}\right| \leq \left|V_j\right|$.

Proof. From the conditions given in the statement of the claim, we can say that there exists $f : V(G_u) \to V(G_k)$ such that $f(v) = \eta_i(v) = \phi(v)$ for all $v \in C_u$ and $\sum_{x\in V(G_u)} \left|\textsc{Decider}_f(x) \cap C_u\right| \leq (\gamma_1 + \frac{\gamma_2-\gamma_1}{2000})n\,|C_u|$

Since $\left|\textsc{Decider}_f(x) \cap C_u\right| = d_H(\mathcal{L}_{C_u}(x), \mathcal{L}_{C_k^i}(f(x)))$, we have

$$\sum_{x\in V(G_u)} d_H(\mathcal{L}_{C_u}(x), \mathcal{L}_{C_k^i}(f(x))) \leq (\gamma_1 + \frac{\gamma_2-\gamma_1}{2000})n\,|C_u|$$

Since we are taking the vertices in W uniformly at random from G_u, we can say that

$$\mathbb{E}\left[\sum_{w\in W} d_H(\mathcal{L}_{C_u}(w), \mathcal{L}_{C_k^i}(f(w)))\right] \leq (\gamma_1 + \frac{\gamma_2-\gamma_1}{2000})\,|C_u|\,|W|$$

Using Hoeffding's inequality, we have

$$\mathbb{P}\left(\sum_{w\in W} d_H(\mathcal{L}_{C_u}(w), \mathcal{L}_{C_k^i}(f(w))) \leq \frac{2\gamma_2}{5}\,|C_u|\,|W|\right) \geq 1 - e^{-\mathcal{O}(|W|)}$$

Now, we define $\psi : W \to \{\ell_1, \ldots, \ell_t\}$ such that $\psi(w) = \mathcal{L}_{C_k^i}(f(w))$. In other words, the C_k^i-labelling of $f(w)$ is same as the labelling of $\psi(w)$ for each $w \in W$. Thus, the ψ defined here satisfies the Condition (i) of this claim, that is, $\sum_{w\in W} d_H(\mathcal{L}_{C_u}(w), \psi(w)) \leq \frac{2\gamma_2}{5}\,|C_u|\,|W|$.

Observe that

$$\left|\{w \in W : \mathcal{L}_{C_k^i}(f(w)) = l_j\}\right| \leq \left|\{v \in V(G_k) : \mathcal{L}_{C_k^i}(v) = l_j\}\right| \leq \left|V_j\right|.$$

So, by the definition of ψ, $\left|\{w \in W : \psi(w) = l_j\}\right| \leq \left|V_j\right|$. Hence ψ considered above also satisfies Condition (ii) of the claim. □

Now consider the situation when the algorithm is at **Step 1** of **Phase 3**. If G_u and G_k are γ_1-close, that is, there exists a bijection ϕ from $V(G_u)$ to $V(G_k)$ such that $d_\phi(G_u, G_k) \leq \gamma_1 n^2$, then there exists $C_u \in \mathcal{C}_u$, $\eta_i : C_u \to V(G_k)$, and ψ satisfying the conditions given in Claims 4.8 and 4.10. However, we do not know ϕ. If we construct, though inefficiently, a bijection ϕ' that is same as ϕ with respect to the same $C_u \in \mathcal{C}_u$, $\eta_i : C_u \to V(G_k)$ and ψ (conditions given in Claims 4.8 and 4.10), then the following claim says that the difference between $d_{\phi'}(G_u, G_k)$ and $d_\phi(G_u, G_k)$ is not too *large*.

Claim 4.11. *Let us assume that $\phi : V(G_u) \to V(G_k)$ be a bijection such that $d_\phi(G_u, G_k) \leq \gamma_1 n^2$, and $(C_u, \eta_i) \in \Gamma$ where $C_u \in \mathcal{C}_u$ and $\eta_i : C_u \to V(G_k)$ be an embedding such that*

- *$\forall\, v \in C_u$ we have $\eta_i(v) = \phi(v)$, and*
- *$\mathrm{EMD}\left(X_{C_u}, Y_{C_k^i}\right) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000}) n \left|C_u\right|$ where $C_k^i = \eta_i(C_u)$.*

Let $\{\ell_1, \ldots, \ell_t\}$ be the all possible C_k^i-labellings of the vertices of G_k where $t = \left\lceil 2^{\left|C_k^i\right|} \right\rceil$, and W be the set of vertices of G_u sampled at random in **Step 1** *of* **Phase 2** *and $\psi : W \to \{\ell_1, \ldots, \ell_t\}$ be the mapping considered in* **Step 2 (iii)** *in* **Phase 2** *such that*

- *$\sum\limits_{w \in W} d_H(\mathcal{L}_{C_u}(w), \psi(w)) \leq \frac{2\gamma_2}{5} \left|C_u\right| \left|W\right|$, and*
- *$\forall j \in [t]$, we have $\left|\{w : \psi(w) = l_j\}\right| \leq \left|V_j\right|$.*

Then, with probability at least $18/21$, there exists a bijection $\phi' : V(G_u) \to V(G_k)$, with $\phi'(x) = \phi(x) = \eta_i(x)$ for each $x \in C_u$ and $\phi'(w) = \hat{\phi}(w)$ for each $w \in W$ such that

$$d_{\phi'}(G_u, G_k) \leq d_\phi(G_u, G_k) + (4\gamma_1 + \frac{\gamma_2 - \gamma_1}{2}) n^2.$$

Proof. We will prove the claim by contradiction. Suppose that

$$d_{\phi'}(G_u, G_k) > d_\phi(G_u, G_k) + (4\gamma_1 + \frac{\gamma_2 - \gamma_1}{2}) n^2 \tag{5}$$

By using Definition A.2, we write the above equation as

$$\sum_{x \in V(G_u)} \left|\textsc{Decider}_{\phi'}(x)\right| > \sum_{x \in V(G_u)} \left|\textsc{Decider}_{\phi}(x)\right| + (4\gamma_1 + \frac{\gamma_2 - \gamma_1}{2}) n^2$$

So,

$$\sum_{x \in V(G_u)} \left|\textsc{Decider}_{\phi'}(x) \Delta \textsc{Decider}_{\phi}(x)\right| > (4\gamma_1 + \frac{\gamma_2 - \gamma_1}{2}) n^2$$

Let us denote $\textsc{Decider}_{\phi'}(x) \Delta \textsc{Decider}_{\phi}(x) = \mathrm{Symm}_{\phi\phi'}(x)$. Dividing the sum in the left hand side with respect to the values of $\left|\textsc{Decider}_{\phi'}(x) \Delta \textsc{Decider}_{\phi}(x)\right|$'s, that is, $\left|\mathrm{Symm}_{\phi\phi'}(x)\right|$'s, we get

$$\sum_{\substack{x \in V(G_u) \\ \left|\mathrm{Symm}_{\phi\phi'}(x)\right| \geq \frac{(\gamma_2 - \gamma_1)n}{1000}}} \left|\mathrm{Symm}_{\phi\phi'}(x)\right| + \sum_{\substack{x \in V(G_u) \\ \left|\mathrm{Symm}_{\phi\phi'}(x)\right| < \frac{(\gamma_2 - \gamma_1)n}{1000}}} \left|\mathrm{Symm}_{\phi\phi'}(x)\right| > (4\gamma_1 + \frac{\gamma_2 - \gamma_1}{2}) n^2$$

Note that the second sum of the left hand side is at most $\frac{\gamma_2 - \gamma_1}{1000} n^2$. Therefore,

$$\sum_{\substack{x \in V(G_u): \\ \left|\mathrm{Symm}_{\phi\phi'}(x)\right| \geq \frac{(\gamma_2 - \gamma_1)n}{1000}}} \left|\mathrm{Symm}_{\phi\phi'}(x)\right| > (4\gamma_1 + \frac{\gamma_2 - \gamma_1}{2}) n^2 - \frac{\gamma_2 - \gamma_1}{1000} n^2 \tag{6}$$

Before proceeding further, consider the following observation, which we will prove in Appendix D.1.

Observation 4.12 ($\star$)**.** If $\left|\mathrm{Symm}_{\phi\phi'}(x)\right| \geq \frac{(\gamma_2-\gamma_1)n}{1000}$, then

$$\mathbb{P}\left(\left|\mathrm{Symm}_{\phi\phi'}(x) \cap C_u\right| \geq (1-\frac{1}{50})\left|\mathrm{Symm}_{\phi\phi'}(x)\right| \frac{|C_u|}{n}\right) \leq e^{-\mathcal{O}(|C_u|)}.$$

This implies that the following holds with probability at least $1 - ne^{-\mathcal{O}(|C_u|)}$.

$$\begin{aligned}\sum_{\substack{x\in V(G_u):\\ \left|\mathrm{Symm}_{\phi\phi'}(x)\right| \geq \frac{(\gamma_2-\gamma_1)n}{1000}}} \left|\mathrm{Symm}_{\phi\phi'}(x) \cap C_u\right| &\geq \left(1-\frac{1}{50}\right)\frac{|C_u|}{n} \sum_{\substack{x\in V(G_u):\\ \left|\mathrm{Symm}_{\phi\phi'}(x)\right| \geq \frac{(\gamma_2-\gamma_1)n}{1000}}} \left|\mathrm{Symm}_{\phi\phi'}(x)\right| \\ &= \frac{49}{50}\left(4\gamma_1 + \frac{499(\gamma_2-\gamma_1)}{1000}\right) n\,|C_u|\,. \ (\because \text{ By Equation 6})\end{aligned}$$

Hence, with probability at least $1 - ne^{-\mathcal{O}(|C_u|)}$, the following event holds.

$$\sum_{x\in V(G_u)} \left|\mathrm{Symm}_{\phi\phi'}(x) \cap C_u\right| \geq \frac{49}{50}\left(4\gamma_1 + \frac{499(\gamma_2-\gamma_1)}{1000}\right) n\,|C_u|\,. \tag{7}$$

Assuming Equation (7) holds and using the fact that $W \subset V(G_u)$ is taken uniformly at random, we can say that

$$\mathbb{E}\left[\sum_{w\in W}\left|\mathrm{Symm}_{\phi\phi'}(x) \cap C_u\right|\right] > \frac{49}{50}(4\gamma_1 + \frac{499(\gamma_2-\gamma_1)}{1000})\,|C_u|\,|W|$$

Using Hoeffding's inequality (See Lemma E.3), we get

$$\mathbb{P}\left(\sum_{w\in W}\left|\mathrm{Symm}_{\phi\phi'}(w) \cap C_u)\right| \leq (3\gamma_1 + \frac{11(\gamma_2-\gamma_1)}{24})\,|C_u|\,|W|\right) \leq e^{-\mathcal{O}(\frac{|C_u|^2|W|^2}{|W||C_u|^2})} = e^{-\mathcal{O}(|W|)}$$

As the above equation holds in the conditional space that Equation (7) holds, we have

$$\mathbb{P}\left(\sum_{w\in W}\left|\mathrm{Symm}_{\phi\phi'} \cap C_u)\right| > (3\gamma_1 + \frac{11(\gamma_2-\gamma_1)}{24})\,|C_u|\,|W|\right) \geq 1 - ne^{-\mathcal{O}(|C_u|)} - e^{-\mathcal{O}(|W|)}. \tag{8}$$

Note that Equation (5) implies Equation (8). However, till now, we have not used any information given in the statement of Claim 4.11, except that C_u and W are taken uniformly at random. By using the fact that the sum of label differences of the vertices of W under C_u-labelling and that of ψ is bounded, we will deduce that

$$\mathbb{P}\left(\sum_{w\in W}\left|\mathrm{Symm}_{\phi\phi'}(w) \cap C_u\right| \leq (2\gamma_1 + \frac{9(\gamma_2-\gamma_1)}{20})\,|C_u|\,|W|\right) \geq 1 - ne^{-\mathcal{O}(|C_u|)} - e^{-\mathcal{O}(|W|)}. \tag{9}$$

As Equation (5) implies Equation (8), and Equations (8) and (9) together implies that Equation (5) does not hold with probability at least $1 - 4ne^{-\mathcal{O}(|C_u|)} - e^{-\mathcal{O}(|W|)}$. Hence, we are done with the proof of Claim 4.11 except that we need to show Equation (9).

By the definition of the bijection ϕ, we have $\sum\limits_{x \in V(G_u)} \left|\textsc{Decider}_\phi(x)\right| \leq \gamma_1 n^2$. This implies

$$\sum_{\substack{x \in V(G_u) \\ \left|\textsc{Decider}_\phi(x)\right| \geq \frac{(\gamma_2-\gamma_1)n}{1000}}} \left|\textsc{Decider}_\phi(x)\right| \leq \gamma_1 n^2 \tag{10}$$

To proceed further, we need the following observation.

Observation 4.13 ($\star$). (i) If $\left|\textsc{Decider}_\phi(x)\right| \geq \frac{(\gamma_2-\gamma_1)n}{1000}$, then

$$\mathbb{P}\left(\left|\textsc{Decider}_\phi(x) \cap C_u\right| \geq (1+\frac{1}{50})\left|(\textsc{Decider}_\phi(x)\right| \frac{|C_u|}{n}\right) \leq e^{-\mathcal{O}(|C_u|)}.$$

(ii) If $\left|\textsc{Decider}_\phi(x)\right| < \frac{(\gamma_2-\gamma_1)n}{1000}$, then $\mathbb{P}\left(\left|\textsc{Decider}_\phi(x) \cap C_u\right| \geq \frac{\gamma_2-\gamma_1}{750}\left|C_u\right|\right) \leq e^{-\mathcal{O}(|C_u|)}$.

The above observation follows from Chernoff bound (See Lemma E.1) and is presented in Appendix D.2, and it implies that the following holds with probability at least $1 - ne^{-\mathcal{O}(|C_u|)}$.

$$\begin{aligned}
&\sum_{x \in V(G_u)} \left|\textsc{Decider}_\phi(x) \cap C_u\right| \\
&= \sum_{\substack{x \in V(G_u): \\ \left|\textsc{Decider}_\phi(x)\right| \geq \frac{(\gamma_2-\gamma_1)n}{1000}}} \left|\textsc{Decider}_\phi(x) \cap C_u\right| + \sum_{\substack{x \in V(G_u): \\ \left|\textsc{Decider}_\phi(x)\right| < \frac{(\gamma_2-\gamma_1)n}{1000}}} \left|\textsc{Decider}_\phi(x) \cap C_u\right| \\
&\leq \left(1+\frac{1}{50}\right) \sum_{\substack{x \in V(G_u): \\ \left|\textsc{Decider}_\phi(x)\right| \geq \frac{(\gamma_2-\gamma_1)n}{1000}}} \left|\textsc{Decider}_\phi(x)\right| \frac{|C_u|}{n} + \frac{(\gamma_2-\gamma_1)n\,|C_u|}{750} \\
&\leq \frac{51}{50}\gamma_1 n\,|C_u| + \frac{(\gamma_2-\gamma_1)n\,|C_u|}{750}
\end{aligned}$$

Note that the last inequality follows from Equation (10). Summarizing the above calculation, we get that the following event occurs with probability at least $1 - ne^{-\mathcal{O}(|C_u|)}$.

$$\sum_{x \in V(G_u)} \left|\textsc{Decider}_\phi(x) \cap C_u\right| \leq \frac{51}{50}\gamma_1 n\,|C_u| + \frac{(\gamma_2-\gamma_1)n\,|C_u|}{750}. \tag{11}$$

Let us assume Equation (11) holds. Since we are taking the vertices of W uniformly at random from $V(G_u)$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{w \in W} \left|\textsc{Decider}_\phi(w) \cap C_u\right|\right] &= \mathbb{E}\left[\sum_{w \in W} d_H(\mathcal{L}_{C_u}(w), \mathcal{L}_{C_k^i}(\phi(w))\right] \\
&\leq \frac{51}{50}\gamma_1\,|C_u|\,|W| + \frac{(\gamma_2-\gamma_1)\,|C_u|\,|W|}{750}.
\end{aligned}$$

Similarly from **Step 2 (iii)** of **Phase 2**, we have

$$
\begin{aligned}
\sum_{w\in W}\left|\textsc{Decider}_{\phi'}(w)\cap C_u\right| &= \sum_{w\in W} d_H(\mathcal{L}_{C_u}(w),\mathcal{L}_{C_k^i}(\phi'(w))) \\
&\le \frac{2\gamma_2}{5}\left|C_u\right|\left|W\right|
\end{aligned}
$$

Recall that $\text{Symm}_{\phi\phi'}(x) = \textsc{Decider}_{\phi'}(x)\Delta\textsc{Decider}_{\phi}(x)$. Therefore,

$$
\begin{aligned}
\mathbb{E}\left[\sum_{w\in W}\left|\text{Symm}_{\phi\phi'}(x)\cap C_u\right|\right] &\le \mathbb{E}\left[\sum_{w\in W}\left|\textsc{Decider}_{\phi'}(w)\cap C_u\right|\right] + \sum_{w\in W}\left|\textsc{Decider}_{\phi}(w)\cap C_u\right| \\
&\le (\frac{764}{750}\gamma_1 + \frac{301(\gamma_2-\gamma_1)}{750})\left|C_u\right|\left|W\right|
\end{aligned}
$$

Using Hoeffding's inequality (see Lemma E.3), we can say that

$$
\mathbb{P}\left(\sum_{w\in W}\left|\text{Symm}_{\phi\phi'}(w)\cap C_u\right| > (2\gamma_1 + \frac{9(\gamma_2-\gamma_1)}{20})\left|C_u\right|\left|W\right|\right) \le e^{-\mathcal{O}(\frac{|C_u|^2|W|^2}{|W||C_u|^2})} = e^{-\mathcal{O}(|W|)}.
$$

Note that the above equation holds on the conditional space that Equation (11) holds. Hence,

$$
\mathbb{P}\left(\sum_{w\in W}\left|\text{Symm}_{\phi\phi'}(w)\cap C_u\right| \le (2\gamma_1 + \frac{9(\gamma_2-\gamma_1)}{20})\left|C_u\right|\left|W\right|\right) \ge 1 - ne^{-\mathcal{O}(|C_u|)} - e^{-\mathcal{O}(|W|)}.
$$

□

If we had constructed a bijection ϕ' as stated in the above claim, we could easily test by sampling *suitable* many random edges from G_u and checking the corresponding edges in G_k. It is important to note that, it is not possible to construct ϕ' efficiently. However, without constructing the bijection ϕ', if we can test for presence of some randomly chosen edges in G_u and their corresponding edges in G_k, we are done. In order to achieve this, we choose W randomly in **Step 1** of **Phase 2** and pair up the vertices of W in **Step 1** of **Phase 3**. Using **Step 2 (iii)** of **Phase 2** and **Step 3** of **Phase 3**, we check if $\hat{\phi}(w) = \phi'(w)$ for each $w \in W$. Note that $\hat{\phi} : W \to V(G_k)$ is the map constructed in **Step 3** of **Phase 3** and $\phi' : V(G_u) \to V(G_k)$ is the bijection as stated in Claim 4.11. Then we check the edge mismatches between the paired up vertices of W in G_u and their corresponding mapped vertices in G_k in **Step 4** of **Phase 3**, which is possible as we have constructed the mappings of the vertices in W in **Step 2 (iii)** of **Phase 2**.

The following claim proves that if G_u and G_k are γ_1-close, then $\zeta(C_u, \eta_i, \psi, \hat{\phi}) \le 5\gamma_1 + \frac{3}{5}(\gamma_2-\gamma_1)$, as considered in **Step 4** of **Phase 3** holds with probability at least $20/21$.

Claim 4.14. *Let us assume that $\phi : V(G_u) \to V(G_k)$ be a bijection such that $d_\phi(G_u, G_k) \le \gamma_1 n^2$, and $(C_u, \eta_i) \in \Gamma$ where $C_u \in \mathcal{C}_u$, and $\eta_i : C_u \to V(G_k)$ be an embedding of C_u such that*

- *$\forall\, v \in C_u$ we have $\eta_i(v) = \phi(v)$, and*
- *$\text{EMD}\left(X_{C_u}, Y_{C_k^i}\right) \le (\gamma_1 + \frac{\gamma_2-\gamma_1}{2000})n\left|C_u\right|$ where $C_k^i = \eta_i(C_u)$.*

Let $\{\ell_1, \ldots, \ell_t\}$ be the all possible C_k^i-labellings of G_k where $t = \left[2^{|C_k^i|}\right]$, W be the set of vertices of G_u sampled at random in **Step 1** *of* **Phase 2***, and $\psi : W \to \{\ell_1, \ldots, \ell_t\}$ be the mapping considered in* **Step 2 (iii)** *of* **Phase 2** *such that*

- $\sum\limits_{w \in W} d_H(\mathcal{L}_{C_u}(w), \psi(w)) \leq \frac{2\gamma_2}{5} |C_u| |W|$*, and*
- $\forall j \in [t]$*, we have* $\left|\{w : \psi(w) = l_j\}\right| \leq \left|V_j\right|$.

If we take an embedding $\hat{\phi} : W \to V(G_k)$ such that $\hat{\phi}(w) \in V_j$ if and only if $\psi(w) = \ell_j$, then

$$\zeta(C_u, \eta_i, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)$$

holds with probability at least $20/21$, where $\zeta(C_u, \eta_i, \psi, \hat{\phi})$ is as defined in **Step 3** *of* **Phase 3***.*

Proof. Recall that W is a subset of $V(G_u)$ taken uniformly at random in **Step 1** of **Phase 2** and we paired up the vertices of W randomly in **Step 1** of **Phase 3** respectively. Also, we are checking the edge mismatches of the paired up vertices of W and their corresponding mapped vertices in G_k according to the mapping $\hat{\phi} : W \to V(G_k)$ in **Step 4** of **Phase 3** to compute $\zeta(C_u, \eta_i, \psi, \hat{\phi})$. Considering the conditions given in the statement of this claim and Claim 4.11, one can think that we are checking the presence of $\frac{|W|}{2}$ many randomly chosen edges in G_u and the corresponding edges in G_k according to some bijection $\phi' : V(G_u) \to V(G_k)$, where ϕ' is a bijection with $d_{\phi'}(G_u, G_k) \leq (5\gamma_1 + \frac{\gamma_2 - \gamma_1}{2})n^2$.

So, $\mathbb{E}\left[\zeta(C_u, \eta_i, \psi, \hat{\phi})\right] \leq (5\gamma_1 + \frac{\gamma_2-\gamma_1}{2})$. Now, applying Hoeffding's inequality (Lemma E.3) and taking $|W| = C' \frac{\log^2 n}{(\gamma_2 - \gamma_1)^3}$ for suitably large constant C', we have

$$\begin{aligned} \mathbb{P}\left(\zeta(C_u, \eta_i, \psi, \hat{\phi}) > 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)\right) &= \mathbb{P}\left(\zeta(C_u, \eta_i, \psi, \hat{\phi})\,|W| > \left(5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)\right)|W|\right) \\ &\leq e^{-\mathcal{O}(|W|)} \leq \frac{1}{21} \end{aligned}$$

□

Now we are ready to prove the completeness property using Claims 4.8, 4.10, 4.11, 4.14 and Theorem 4.2.

Lemma 4.15 (Completeness Lemma)**.** *If G_u and G_k are γ_1-close to isomorphic, then our algorithm reports the same with probability at least $2/3$.*

Proof. Observe that from Claim 4.8, we know that, with probability at least $20/21$, there exists a $C_u \in \mathcal{C}_u$ and an embedding $\eta_i : C_u \to V(G_k)$ such that $\text{EMD}\left(X_{C_u}, Y_{C_k^i}\right) \leq \left(\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000}\right) n\,|C_u|$ where $C_k^i = \eta_i(C_u)$. Similarly, from Theorem 4.2, we can say that, with probability at least $20/21$, the algorithm ALG-EMD returns all embeddings η_i such that $\text{EMD}\left(X_{C_u}, Y_{C_k^i}\right) \leq \left(\gamma_1 + \frac{\gamma_2-\gamma_1}{2000}\right) n\,|C_u|$. Now from Claim 4.10, we know that, with probability at least $20/21$, conditions of Equation (3) hold. Again, from Claim 4.11, we can say that constructing partial bijection at **Step 3** of **Phase 3** does not change isomorphism distance by more than $(4\gamma_1 + \frac{\gamma_2 - \gamma_1}{2})n^2$ with probability at least $18/21$.

Finally, from Claim 4.14, we can say that the algorithm will correctly detect the distance at **Step 4** of **Phase 3** by testing $\zeta(C_u, \eta_i, \psi, \hat{\phi}) \le 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)$ with probability at least 20/21. Thus, using union bound, we can say that when G_k and G_u are γ_1-close to being isomorphic, **TolerantGI(**G_u**,** G_k**,** γ_1**,** γ_2**)** reports the same with probability at least 2/3. □

4.2.2 Proof of Soundness Property

Similarly for the soundness property of our algorithm, let us consider the case when G_u and G_k are γ_2-far from being isomorphic. Then we will show that the algorithm will output the correct answer with probability at least 2/3.

Recall the definition of the set Γ_W with which we started **Phase 3** of our algorithm.

$$\Gamma_W = \{(C_u, \eta_i, \psi) : (C_u, \eta_i) \in \Gamma \text{ such that Equation 3 holds}\}.$$

By Observation 4.5, we have

$$\Pr\left(\forall\ (C_u, \eta_i, \psi) \in \Gamma_W, EMD(X_{C_u}, Y_{C_k^i}) \le \frac{\gamma_2}{5}|C_u|n\right) \ge \frac{8}{9}. \tag{12}$$

From now on, we work on the conditional space where $EMD(X_{C_u}, Y_{C_k^i}) \le \frac{\gamma_2}{5}|C_u|n\ \forall\ (C_u, \eta_i, \psi)$ holds. By Observation 4.7 (i), we know that $|\Gamma_W| \le 2^{\mathcal{O}(\log^2 n/(\gamma_2-\gamma_1))}$. So, the following claim about any $(C_u, \eta_i, \psi) \in \Gamma_W$ along with union bound over all the elements in Γ_W, we will be done with the proof of soundness property.

Claim 4.16. *Let $(C_u, \eta_i, \psi) \in \Gamma_W$ and $\hat{\phi}$ be the embedding of W into G_k constructed while executing* **Step 3** *of* **Phase 3** *for (C_u, η_i, ψ). Also, let $EMD(X_{C_u}, Y_{C_k^i}) \le \frac{\gamma_2}{5}|C_u|n$, where $C_k^i = \eta_i(C_u)$. Then the following holds with probability at most $\frac{2}{9|\Gamma_W|}$:*

$$\zeta(C_u, \eta_i, \psi, \hat{\phi}) \le 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1).$$

Proof. Let $\Phi(C_u, C_k^i)$ be the class of all bijections such that the following hold for each $\phi \in \Phi(C_u, C_k)$.

- $\phi(x) = \eta_i(x)$ for each $x \in C_u$, and
- $\sum\limits_{v \in V(G_u)} \left|\text{DECIDER}_\phi(v) \cap C_u\right| \le \frac{\gamma_2}{5} n\,|C_u|$.

Consider the following observation, about the bijections in Φ, that we will prove later.

Observation 4.17. Let ϕ be a bijection in Φ. Then $\sum\limits_{w \in W} \left|\text{DECIDER}_\phi(w) \cap C_u\right| \le \frac{2\gamma_2}{5}\,|C_u|\,|W|$ holds with probability at least $1 - \frac{1}{9|\Gamma_W|}$.

Our algorithm constructs $\psi : W \to \{\ell_1, \dots, \ell_t\}$ in **Step 2** of **Phase 2** satisfying

- $\sum\limits_{w \in W} d_H(\mathcal{L}_{C_u}(w), \psi(w)) \le \frac{2\gamma_2}{5}\,|C_u|\,|W|$, and
- $\forall j \in [t]$, we have $\left|\{w : \psi(w) = l_j\}\right| \le \left|V_j\right|$.

Note that $\sum_{w \in W} d_H(\mathcal{L}_{C_u}(w), \psi(w)) = \sum_{w \in W} \left|\text{DECIDER}_\phi(w) \cap C_u\right|$, where ϕ is some bijection in Φ. After getting ψ, we construct a partial bijection $\hat{\phi} : W \to V(G_k)$ that satisfies the above two conditions. So, one can think of W is taken uniformly at random from the set of all W's satisfying $\sum_{w \in W} \left|\text{DECIDER}_\phi(w) \cap C_u\right| \leq \frac{2\gamma_2}{5} |C_u| |W|$. Now, from Observation 4.17, we have the following observation.

Observation 4.18. $\hat{\phi}$ is a *random restriction* of a random bijection $\phi \in \Phi(C_u, C_k)$ by the set W with probability at least $1 - \frac{1}{9|\Gamma_W|}$.

Proof. Let us consider a ϕ such that $\phi|_W = \hat{\phi}$. Let $\mathcal{W} = \{\hat{\phi}_X = \phi|_X : X \subset V(G_u) \text{ and } |X| = |W|\}$, and $\mathcal{W}' \subseteq \mathcal{W}$ is defined as:

$$\mathcal{W}' = \left\{ \hat{\phi}_X \in \mathcal{W} : \sum_{w \in X} \left|\text{DECIDER}_\phi(w) \cap C_u\right| \leq \frac{2\gamma_2}{5} |C_u| |W| \right\}$$

Observe that $\hat{\phi} = \hat{\phi}_W \in \mathcal{W}$. By Observation 4.17, we know that if we take a set $X \subset V(G_u)$ (i.e, a $\hat{\phi}_X$ uniformly at random from $\mathcal{W}$), then the probability that $\hat{\phi}_X \in \mathcal{W}'$, is at least $1 - \frac{1}{9|\Gamma_W|}$. So, $|\mathcal{W}'| \geq \left(1 - \frac{1}{9|\Gamma_W|}\right) |W|$.

Observe that the partial bijection $\hat{\phi}$, constructed by our algorithm, is same as that of $\hat{\phi}_W$, and $\hat{\phi}$ is in $\mathcal{W}'$. Now, using the fact that $|\mathcal{W}'| \geq \left(1 - \frac{1}{9|\Gamma_W|}\right) |W|$, the observation follows. $\square$

Recall that W is a subset of $V(G_u)$ taken uniformly at random in **Step 1** of **Phase 2** and we paired up the vertices of W randomly in **Step 1** of **Phase 3** respectively. Also, we are checking the edge mismatches of the paired up vertices of W and their corresponding mapped vertices in G_k according to the mapping $\hat{\phi} : W \to V(G_k)$ in **Step 4** of **Phase 3** to compute $\zeta(C_u, \eta_i, \psi, \hat{\phi})$. Considering the discussion here, one can think of that, we are checking the presence of $\frac{|W|}{2}$ many randomly chosen edges in G_u and the corresponding edges in G_k according to some bijection $\phi \in \Phi$.

Note that $d_\phi(G_u, G_k) \geq \gamma_2 n^2$. Thus, $\mathbb{E}\left[\zeta(C_u, \eta_i, \psi, \hat{\phi})\right] \geq \gamma_2 |W|$. Now we can deduce the following. [11]

$$\begin{aligned}
\mathbb{P}\left(\zeta(C_u, \eta_i, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)\right) &= \mathbb{P}\left(\zeta(C_u, \eta_i, \psi, \hat{\phi}) |W| \leq (5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)) |W|\right) \\
&\leq e^{-\mathcal{O}(|W|)} \\
&\leq \frac{1}{9 |\Gamma_W|}
\end{aligned}$$

Note that we were deriving the above bound on $\mathbb{P}\left(\zeta(C_u, \eta_i, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)\right)$ assuming that $\hat{\phi}$ is a random restriction of a random $\phi \in \Phi$. Hence, combining Observation 4.18 with the above bound on $\mathbb{P}\left(\zeta(C_u, \eta_i, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)\right)$ (when $\hat{\phi}$ is a random restriction of a random $\phi \in \Phi$), we get

$$\mathbb{P}\left(\zeta(C_u, \eta_i, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)\right) \leq \frac{2}{9 |\Gamma_W|}.$$

[11] Here we are assuming $\gamma_2 \geq 11\gamma_1$.

□

Proof of Observation 4.17. . Since W is taken uniformly at random,

$$\mathbb{E}\left[\sum_{w\in W}\left|\textsc{Decider}_{\phi}(w)\cap C_u\right|\right]\leq\frac{\gamma_2}{5}\left|C_u\right|\left|W\right|$$

Using Hoeffding's inequality, we get

$$\mathbb{P}\left(\sum_{w\in W}\left|\textsc{Decider}_{\phi}(w)\cap C_u\right|\geq\frac{2\gamma_2}{5}\left|C_u\right|\left|W\right|\right)\leq e^{-\mathcal{O}(|W|)}\leq\frac{1}{9\left|\Gamma_W\right|}.$$

□

Now we are ready to prove the soundness property of our algorithm.

Lemma 4.19 (Soundness Lemma)**.** *If G_u and G_k are γ_2-far from isomorphic, then the algorithm reports the same with probability at least* $2/3$*.*

Proof. From Observation 4.7 (i), we know that $|\Gamma_W|$ is at most $2^{C_1\frac{\log^2 n}{\gamma_2-\gamma_1}}$. In Claim 4.16, we are proving that $\zeta(C_u,\eta_i,\psi,\hat{\phi})\leq 5\gamma_1+\frac{3}{5}(\gamma_2-\gamma_1)$ holds with probability at most $\frac{2}{9|\Gamma_W|}$ for any particular $(C_u,\eta_i,\psi)\in\Gamma_W$ with $EMD(X_{C_u},Y_{C_k^i})\leq\frac{\gamma_2}{5}|C_u|n$. So, by the union bound, the probability that there exists a $(C_u,\eta_i,\psi)\in\Gamma_W$ with $EMD(X_{C_u},Y_{C_k^i})\leq\frac{\gamma_2}{5}|C_u|n$ such that $\zeta(C_u,\eta_i,\psi,\hat{\phi})\leq 5\gamma_1+\frac{3}{5}(\gamma_2-\gamma_1)$, is at most $\frac{2}{9}$. Now From Equation 12,

$$\Pr\left(\forall\ (C_u,\eta_i,\psi,\hat{\phi})\in\Gamma_W, EMD(X_{C_u},Y_{C_k^i})\leq\frac{\gamma_2}{5}|C_u|n\right)\geq\frac{8}{9}$$

Putting everything together, the probability that the algorithm reports that G_u and G_k are γ_2-far, is at least $2/3$. □

Till now we have proved the completeness and soundness property of our algorithm **TolerantGI**. We will prove the query complexity property in the next section when we prove the final theorem.

4.3 Proof of Theorem 4.1

Proof. From the *Completeness Lemma* (Lemma 4.15) and *Soundness Lemma* (Lemma 4.19), we can say that our algorithm **TolerantGI** correctly decides whether $d(G_u,G_k)\leq\gamma_1 n^2$ or $d(G_u,G_k)\geq\gamma_2 n^2$ with probability at least $2/3$.

Now, we calculate the query complexity of our algorithm. Note that **Step 1** and **Step 2** of **Phase 1**, **Step 1** and **Step 3** of **Phase 2**, **Step 1**, **Step 2** and **Step 3** of **Phase 3**, of the algorithm **TolerantGI**, do not require any query to the adjacency matrix of G_u. Let $\textsc{Cost}_{C_u}$ denote the query complexity corresponding to a particular $C_u\in\mathcal{C}_u$. So, the total query complexity of the algorithm **TolerantGI** is $\sum\limits_{C_u\in\mathcal{C}_u}\textsc{Cost}_{C_u}$. Observe that

$$\textsc{Cost}_{C_u}\quad=\quad\text{Query Complexity of algorithm ALG-EMD }+\textsc{Cost}_{C_u,W}$$

where $\textsc{Cost}_{C_u,W}$ denotes the query complexity of **Step 1** of **Phase 2** corresponding to W and $C_u \in \mathcal{C}_u$.

Note that ALG-EMD is the algorithm corresponding to Theorem 4.2. In **Step 3** of **Phase 1** of our algorithm, for each $C_u \in \mathcal{C}_u$, we call ALG-EMD with parameters $d = \mathcal{O}\left(\frac{\log n}{\gamma_2-\gamma_1}\right)$, $t = 2^{\mathcal{O}\left(\frac{\log^2 n}{\gamma_2-\gamma_1}\right)}$, $\epsilon_1 = \left(\gamma_1 + \frac{\gamma_2-\gamma_1}{2000}\right)$, $\epsilon_2 = \frac{\gamma_2}{5}$ and $\delta = \Theta(1)$. So, the query complexity of each call, to ALG-EMD from our algorithm, is $\widetilde{\mathcal{O}}\left(\min\{n, 2^d\}\right) = \widetilde{\mathcal{O}}(n)$.

Further note that, from the description **Step 1** of **Phase 2**, $\mathrm{COST}_{C_u,W} = \mathcal{O}\left(\frac{\log^2 n}{\gamma_2-\gamma_1}\right)$. Since $|\mathcal{C}_u| = \mathcal{O}\left(\frac{1}{\gamma_2-\gamma_1}\right)$, the total query complexity of our algorithm is $\widetilde{\mathcal{O}}(n)$. $\square$

5 Conclusion

In this paper, we proved that the query complexity of tolerant GI testing between a known graph G_k and an unknown graph G_u is the same as (up to polylogarithmic factor) testing of EMD between a known multi-set S_k and an unknown multi-set S_u when we have samples **without** replacement from S_u. In Lemma B.10, we have shown that the sample complexity of testing of EMD between a known multi-set S_k and an unknown multi-set S_u when we have samples **with** replacement from S_u is $\Omega(n/\log n)$. Thus the natural open question is

*What is the query complexity of tolerant EMD testing when we have samples **without** replacement from the unknown multi-set?*

It is also interesting to note that our lower bound proof is via a pure reduction from graph isomorphism to testing EMD of multi-sets over the Hamming cube using samples **without** replacement. Using our lower bound technique (and Proposition B.7), we can get an alternative proof of Fischer and Matsliah's lower bound result for testing non-tolerant graph isomorphism [FM08]. Our upper bound proof is also a pure reduction from testing EMD of multi-sets over the Hamming cube to tolerant graph isomorphism problem. Thus our reductions also hold for other computational models such as the communication complexity model. So, in the communication model (that is, when Alice and Bob have graphs G_A and G_B respectively and they want to estimate the GI-distance between them), the amount of bits of communication is same (up to a polylogarithmic factors) to the problem of estimating the EMD distance between two distributions over Hamming cube, where Alice and Bob have access to one distribution each. The question we would like to pose is:

What is the randomized communication complexity of testing tolerant graph isomorphism problem?

Fischer and Matsliah [FM08] studied the non-tolerant version of the graph isomorphism problem in two scenarios: (i) one graph is known and the other graph is unknown, (ii) both the graphs are unknown. They resolved the query complexity of (i), whereas Onak and Sun [OS18] resolved (ii). With this paper, we initiate the study of tolerant graph isomorphism problem in the query world and settled the question completely when one graph is unknown, and the other graph is known. So, another natural open question to look for is:

What is the query complexity of tolerant graph isomorphism when both the graphs are unknown?

6 Acknowledgement

The authors would like to thank the anonymous reviewer for pointing out a mistake in an earlier version of this paper.

References

[Bab16] László Babai. Graph Isomorphism in Quasipolynomial Time. In *Proceedings of the 48th Annual ACM symposium on Theory of Computing, STOC*, pages 684–697, 2016.

[BC10] Laszlo Babai and Sourav Chakraborty. Property Testing of Equivalence under a Permutation Group Action. *ACM Transactions on Computation Theory (ToCT)*, 2010.

[BDST15] László Babai, Anuj Dawar, Pascal Schweitzer, and Jacobo Torán. The Graph Isomorphism Problem (Dagstuhl Seminar 15511). *Dagstuhl Reports*, 5(12):1–17, 2015.

[BFF+01] Tugkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing Random Variables for Independence and Identity. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science, FOCS*, pages 442–451, 2001.

[Can15] Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 22, pages 1–1, 2015.

[DBNNR11] Khanh Do Ba, Huy L Nguyen, Huy N Nguyen, and Ronitt Rubinfeld. Sublinear time algorithms for earth mover's distance. *Theory of Computing Systems*, 48(2):428–442, 2011.

[DL12] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

[DP09] Devdatt P Dubhashi and Alessandro Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[FM08] Eldar Fischer and Arie Matsliah. Testing Graph Isomorphism. *SIAM Journal on Computing*, 38(1):207–225, 2008.

[Fre77] David Freedman. A remark on the difference between sampling with and without replacement. *Journal of the American Statistical Association*, 72(359):681–681, 1977.

[Lin94] Chih-Long Lin. Hardness of Approximating Graph Transformation Problem. In *Proceedings of the 5th International Symposium on Algorithms and Computation, ISAAC,*, pages 74–82, 1994.

[OS18] Krzysztof Onak and Xiaorui Sun. The Query Complexity of Graph Isomorphism: Bypassing Distribution Testing Lower Bounds. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 165–171, 2018.

[PRR06] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *Journal of Computer and System Sciences*, 72(6):1012–1042, 2006.

[Sun16] Xiaorui Sun. *On the Isomorphism Testing of Graphs*. PhD thesis, Columbia University, 2016.

[VV11] Gregory Valiant and Paul Valiant. The Power of Linear Estimators. In *Proceedings of the 52nd IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 403–412, 2011.

A Preliminaries

All graphs considered here are undirected, unweighted and have no self-loops or parallel edges. For a graph $G(V,E)$, $V(G)$ and $E(G)$ will denote the vertex set and the edge set of G, respectively. Since we are considering undirected graphs, we write an edge $(u,v) \in E(G)$ as $\{u,v\}$. The *Hamming distance* between two points x and y in a Hamming cube $\{0,1\}^k$ will be denoted by $d_H(x,y)$.

A.1 Notion of distance between two graphs

First let us define the notion of DECIDER of a vertex and then the notion of distance between two graphs, using decider of vertices, that is conceptually same as that of GRAPH ISOMORPHISM DISTANCE defined in Definition 1.1.

Definition A.1. (DECIDER of a vertex) Given two graphs G_k and G_u and a bijection $\phi : V(G_u) \to V(G_k)$, DECIDER of a vertex $x \in V(G_u)$ with respect to ϕ is defined as the set of vertices of G_u that create the edge difference in x and $\phi(x)$'s neighbourhood in G_u and G_k, respectively. Formally,

$$\textsc{Decider}_{\phi}(x) \quad := \quad \{y \in V(G_u) : \text{ one of the edges } \{x,y\} \textit{ and } \{\phi(x),\phi(y)\} \text{ is not present}\}$$

Definition A.2. (DISTANCE between two graphs) Let G_u and G_k be two graphs and $\phi : V(G_u) \to V(G_k)$ be a bijection from the vertex set of G_u to that of G_k. The *distance* between G_u and G_k under ϕ is defined as the sum of the sizes of the deciders of all the vertices in G_u, that is,

$$d_{\phi}(G_u, G_k) := \sum_{x \in V(G_u)} \left|\textsc{Decider}_{\phi}(x)\right|.$$

The *distance* between two graphs G_u and G_k is the minimum distance under all possible bijections ϕ from $V(G_u)$ to $V(G_k)$, that is, $d(G_u, G_k) := \min_{\phi} d_{\phi}(G_u, G_k)$.

Remark 5. Recall the definition of $\delta_{GI}(G_u, G_k)$, GRAPH ISOMORPHISM DISTANCE between G_u and G_k, that is given in Definition 1.1. Observe that $d(G_u, G_k) = 2\binom{n}{2}\delta_{GI}(G_u, G_k)$. Though, $d(G_u, G_k)$ and $\delta_{GI}(G_u, G_k)$ represent the same thing, conceptually, we will do our calculations by using $d(G_u, G_k)$ for simplicity of presentation.

Next we define the concept of closeness between two graphs.

Definition A.3. (CLOSE and FAR) For $\gamma \in [0,1)$, two graphs G_u and G_k with n vertices are *γ-close to isomorphic* if $d(G_u, G_k) \leq \gamma n^2$. Otherwise, we say G_u and G_k are *γ-far from being isomorphic*. [12]

A.2 Property Testing of Distribution Properties

Understanding different properties of probability distributions have been an active area of research in property testing (For reference, see [Can15]). The authors studied these problems assuming random sample access from the unknown distributions. Considering the relation between the distributions and their corresponding representative multi-sets, we can say that all these results hold for multi-sets along with access over sampling **with** replacement.

Although it seems that the change of query model from sample **with** replacement to sample **without** replacement does not make much difference, following the work of Freedman [Fre77], we know that the variation distance between probability distributions when accessed via samples **with** and **without** replacement, becomes arbitrary close to $1/2$ when the number of samples is $\Omega(\sqrt{n})$. Because of this reason, many techniques developed for sampling **with** replacement for various problems no longer work anymore. Most importantly, proving any lower bound better than $\Omega(\sqrt{n})$ is often nontrivial.

B Earth Mover's Distance (EMD) over Hamming Cube

In this section, we study some properties of *Earth Mover's* distance (*EMD*) over probability distributions and multi-sets, which are crucial in the context of both our lower and upper bound. Before proceeding to the discussion on EMD, let us first recall the definition of ℓ_1 distance between two distributions.

Definition B.1 (ℓ_1 distance between two distributions)**.** Let p and q be two probability distributions over $[n]$. The ℓ_1 distance between p and q is defined as

$$d_{l_1}(p,q) = \sum_{i=1}^{n} |p(i) - q(i)|$$

Definition B.2 (*EMD* between two probability distributions)**.** Let $H = \{0,1\}^d$ be a Hamming cube of dimension d, and p, q be two probability distributions on H. The *EMD* between p and q is denoted by $EMD(p,q)$ and defined as the optimum solution to the following linear program:

$$\begin{aligned} &\text{Minimize} \quad \sum_{x,y \in H} f_{xy} d_H(x,y) \\ &\text{Subject to} \quad \sum_{y \in H} f_{xy} = p(x) \ \forall x \in H, \text{ and } \sum_{x \in H} f_{xy} = q(y) \ \forall y \in H. \end{aligned}$$

Now we define *EMD* between two multi-sets.

Definition B.3 (*EMD* between two multi-sets)**.** Let S_1, S_2 be two multi-sets on a Hamming cube $H = \{0,1\}^d$ of dimension d with $|S_1| = |S_2|$. The *EMD* between S_1 and S_2 is denoted by $EMD(S_1, S_2)$ and defined as $EMD(S_1, S_2) = \min\limits_{\phi: S_1 \to S_2} \sum\limits_{x \in S_1} d_H(x, \phi(x))$ where ϕ is a bijection from S_1 to S_2.

[12] By abuse of notation, we will say G_u and G_k are γ-far when $d(G_u, G_k) \geq \gamma n^2$.

Note that an unknown distribution p is accessed by taking samples from p. However, a multi-set is accessed as follows:

Definition B.4 (Query accesses to multi-sets)**.** A multi-set S of n elements is accessed in one of the following ways:

Sample Access with replacement: Each element of S is reported uniformly at random independent of all previous queries.

Sample Access without replacement: Let us assume we make Q queries to S, where $Q \leq n$. The answer to the first query, say s_1, is an element from S chosen uniformly at random. For any $2 \leq i \leq Q$, the answer of the i-th query is an element chosen uniformly at random from $S \setminus \{s_1, \ldots, s_{i-1}\}$. Here $s_j, 1 \leq j \leq Q$, denotes the answer to the j-th query.

Although sampling **with** replacement is more natural query model, we need sampling **without** replacement for our lower bound proof. We now show that we can simulate samples **with** replacement when we have samples **without** replacement.

Proposition B.5 (Simulating samples **with** replacement from samples **without** replacement)**.** *Given Q many samples **without** replacement from an unknown multi-set S_u with n elements, we can simulate Q many samples **with** replacement from S_u where $Q \leq n$.*

Proof. Consider the following procedure to get Q many samples **with** replacement (say $x_1, \ldots, x_Q$) when we have Q many samples **without** replacement ($s_1, \ldots, s_Q$) from the unknown multi-set S_u with $Q \leq n$.

We first set $x_1 = s_1$. For each i with $2 \leq i \leq Q$, we set x_i as follows: with probability $1 - \frac{i-1}{n}$, we select one of the element from $\{s_1, \ldots, s_{i-1}\}$ uniformly at random as x_i; with probability $\frac{i-1}{n}$, we set $x_i = s_i$. From the description of procedure to generate x_i's, we have $\mathbb{P}\left(x_i = s_i\right) = \frac{1}{n}$.

Thus we can simulate Q many samples **with** replacement from Q many samples **without** replacement from the unknown multi-set S_u. □

The following observation connects the EMD between two probability distributions with that of between two multi-sets.

Observation B.6. Let p, q be two probability distributions, having support size K, on a n dimensional Hamming cube $H = \{0,1\}^n$. Then p and q induces two multi-sets S_1 and S_2 on H, respectively, as follows. S_1 (S_2) is the multi-set containing $x \in H$ with multiplicity $p(x)K$ ($q(x)K$) for each $x \in H$. Moreover, $EMD(p, q) = \frac{EMD(S_1, S_2)}{K}$.

Proof. Recall the definitions of EMD between two distributions and two multi-sets given in Definition B.2 and B.3, respectively. We will be done with the proof by showing $EMD(S_1, S_2) \leq K \cdot EMD(p, q)$ and $K \cdot EMD(p, q) \leq EMD(S_1, S_2)$, separately.

For $EMD(S_1, S_2) \leq K \cdot EMD(p, q)$, let $\{f^*_{ij} : i, j \in H\}$ be the set of variables that realizes $EMD(p, q)$, that is, $EMD(p, q) = \sum\limits_{i,j \in H} f^*_{ij} d_H(i, j)$. Consider a bijection ϕ from S_1 to S_2 where $\phi(i) = j$ for g_{ij} many i's. Hence, by Definition B.3,

$$EMD(S_1, S_2) \leq \sum_{x \in S_1} d_H(x, \phi(x)) = \sum_{i,j \in H} K \cdot f^*_{ij} d_H(i, j) = K \cdot EMD(p, q).$$

Now, we show $K \cdot EMD(p,q) \leq EMD(S_1, S_2)$. Let ϕ^* be a bijection from S_1 to S_2 that realizes $EMD(S_1, S_2)$, that is, $EMD(S_1, S_2) = \sum_{x \in S_1} d_H(x, \phi^*(x))$. For any $x, y \in H$, let f_{xy} be the number of elements, of the form (x, y) in $S_1 \times S_2$ such that x is mapped to y under ϕ, divided by K^2. Observe that $f_{xy} \geq 0$. Also, $f_{xy} > 0$ if and only if $(x, y) \in S_1 \times S_2$. More over, $\{f_{ij} : i, j \in H\}$ satisfies $\sum_{i \in H} f_{ij} = p(j)\ \forall j \in H$ and $\sum_{j \in H} f_{ij} = q(i)\ \forall i \in H$. Hence, by Definition B.2,

$$
\begin{aligned}
K \cdot EMD(p,q) \leq K \sum_{x,y \in H} f_{xy} d_H(x,y) &= \sum_{(x,y) \in S_1 \times S_2} K \cdot f_{xy} d_H(x,y) \\
&= \sum_{x \in S_1} d_H(x, \phi^*(x)) = EMD(S_1, S_2).
\end{aligned}
$$

□

Remark 6. Note that sample access from a probability distribution is exactly same as uniform sampling from a multi-set **with** replacement.

Proposition B.7. *Let $\mathcal{D}$ be the set of all multi-sets of size n over a universe $[m]$; let S_k and S_u in $\mathcal{D}$ denote the known and unknown multi-sets over $[n]$; and* PROP : $\mathcal{D} \times \mathcal{D} \to \{0,1\}$ *be a boolean function. Then the following holds:*

If there exists an algorithm that determines PROP *by Q many samples **without** replacement from S_u with probability at least $2/3$, then there exists an algorithm that determines* PROP *by $\min\{Q, \sqrt{\min\{n,m\}}\}$ many samples **with** replacement from S_u with probability at least $2/3 - o(1)$.*

This follows from the fact that when $Q = o(\sqrt{n})$ and D_{WR} (D_{WoR}) be the probability distribution over all the subsets having Q elements from $[n]$ **with** (**without**) replacement, the ℓ_1 distance between D_{WR} and D_{WoR} is $o(1)$.

Definition B.8 (EMD over multi-sets while sampling with and without replacement). Let S_k and S_u denote the known and the unknown multi-sets, respectively, over n-dimensional Hamming cube $H = \{0,1\}^n$ such that $|S_u| = |S_k| = n$. Consider the two distributions p_u and p_k over the Hamming cube H that are naturally defined by the sets S_u and S_k where for all $x \in H$ probability of x in p_u (and p_k) is the number of occurrences of x in S_u (and S_k) divided by n. We then define the EMD between the multi-sets S_u and S_k as

$$EMD(S_u, S_k) \triangleq n \cdot EMD(p_u, p_k).$$

The problem of estimating the EMD over multi-sets while sampling **with** (or **without**) replacement means designing an algorithm, that given any two constants β_1, β_2 such that $0 \leq \beta_1 < \beta_2 \leq 1$, and access to the unknown set S_u by sampling **with** (or **without**) replacement decides whether $EMD(S_k, S_u) \leq \beta_1 n^2$ or $EMD(S_k, S_u) \geq \beta_2 n^2$ with probability at least $2/3$.

Note that estimating the EMD over multi-sets while sampling **with** replacement is exactly same as estimating EMD between the distributions p_u and p_k with samples drawn according to p_u.

Let $\text{QWR}_{\text{EMD}}(n, d, \beta_1, \beta_2)$ ($\text{QWoR}_{\text{EMD}}(n, d, \beta_1, \beta_2)$) denote the number of samples **with** (**without**) replacement required to decide the above from the unknown multi-set S_u. For ease of presentation, we write $\text{QWoR}_{\text{EMD}}(n, d)$ ($\text{QWR}_{\text{EMD}}(n, d)$) instead of $\text{QWoR}_{\text{EMD}}(n, d)$ ($\text{QWR}_{\text{EMD}}(n, \beta_1, \beta_2)$) when the proximity parameters are clear from the context.

Proposition B.9 (Query complexity of EMD increases with number of points as well as dimension). *Let $n, n_1, n_2, d, d_1, d_2 \in \mathbb{N}$ be such that $d_1 \leq d_2$ and $n_1 \leq n_2$. Then*

(i) $\text{QWR}_{\text{EMD}}(n_1, d) \leq \text{QWR}_{\text{EMD}}(n_2, d)$ *and* $\text{QWoR}_{\text{EMD}}(n_1, d) \leq \text{QWoR}_{\text{EMD}}(n_2, d)$;

(ii) $\text{QWR}_{\text{EMD}}(n, d_1) \leq \text{QWR}_{\text{EMD}}(n, d_2)$ *and* $\text{QWoR}_{\text{EMD}}(n, d_1) \leq \text{QWoR}_{\text{EMD}}(n, d_2)$.

Remark 7. For $d = n$ (as considered in Definition 1.2), $\text{QWoR}_{\text{EMD}}(n, d)$ ($\text{QWR}_{\text{EMD}}(n, d)$) are denoted as $\text{QWoR}_{\text{EMD}}(n)$ ($\text{QWR}_{\text{EMD}}(n)$).

Now let us state the lower bound of $\text{QWR}_{\text{EMD}}(n)$.

Theorem B.10. $\text{QWR}_{\text{EMD}}(n) = \Omega(\frac{n}{\log n})$.

Thus following Proposition B.7, we have

Theorem B.11. $\text{QWoR}_{\text{EMD}}(n) = \Omega(\sqrt{n})$.

Note that an upper bound of $\text{QWoR}_{\text{EMD}}(n) = \widetilde{\mathcal{O}}(n)$ is trivial. In the rest of the section, we focus on proving Theorem B.10 that states the lower bound on $\text{QWR}_{\text{EMD}}(n)$. We also provide an upper bound for $\text{QWR}_{\text{EMD}}(n)$ at Lemma B.16 that shows that $\widetilde{\mathcal{O}}(n)$ many samples **with** replacement from S_u to estimate $\text{QWR}_{\text{EMD}}(n)$. Note that by Remark 6, it is enough to show the following lemma that states the lower bound for tolerant EMD testing between two distributions.

Lemma B.12. *Let p and q be two known and unknown distributions, respectively, supported over a subset S of a Hamming cube $H = \{0,1\}^n$ with $|S| = n$. Then there exists a constant ϵ_{EMD} such that the following holds. Given two constants β_1, β_2 with $0 < \beta_1 < \beta_2 < \epsilon_{EMD}(c)$, $\Omega\left(\frac{n}{\log n}\right)$ samples from the distribution q are necessary in order to decide whether $EMD(p, q) \leq \beta_1 n$ or $EMD(p, q) \geq \beta_2 n$. Moreover, $\epsilon_{EMD} = \frac{1-\epsilon_{\ell_1}}{4}$, where ϵ_{ℓ_1} is the constant that is mentioned in Theorem B.14.*

To prove the above lower bound, let us first consider the following lower bound for tolerant ℓ_1 testing between two probability distributions.

Theorem B.13 (Valiant and Valiant [VV11]). *Let p and q be two known and unknown probability distributions respectively over $[n]$. There is an absolute constant ϵ such that in order to decide whether $\|p - q\|_1 \leq \epsilon$ or $\|p - q\|_1 \geq 1 - \epsilon$, $\Omega(\frac{n}{\log n})$ samples, from the distribution q, are necessary.* [13]

Now, we restate the above result for our purpose.

Theorem B.14. *Let p and q be two known and unknown probability distributions, having support size n, over a Hamming cube $H = \{0,1\}^n$. There is an absolute constant ϵ_{ℓ_1} such that in order to decide whether $\|p - q\|_1 \leq \alpha_1$ or $\|p - q\|_1 \geq \alpha_2$ with $0 < \alpha_1 < \alpha_2 \leq 1 - \epsilon_{\ell_1}$, $\Omega(\frac{n}{\log n})$ samples, from the distribution q, are necessary.*

As noted earlier, we will prove Theorem B.10 by using Lemma B.14. However, Theorem B.10 is regarding *EMD* between two distributions whereas Lemma B.14 is regarding ℓ_1 distance between two distributions. The following observation (from [DBNNR11]) gives a connection between *EMD* between two distributions with the ℓ_1 distance between them, which will be required in lower bound proof.

[13] Note that this is rephrasing of the result proved in [VV11]. For reference, see Chapter 3 of the survey by Canonne [Can15].

Proposition B.15 ([DBNNR11]). *Let (M, D) be a finite metric space and p and q be two probability distributions on M. Minimum distance between any two points of M is $\Delta_{\min}$ and diameter of M is $\Delta_{\max}$. Then the following condition holds:*

$$\frac{\|p-q\|_1 \Delta_{\min}}{2} \leq EMD(p,q) \leq \frac{\|p-q\|_1 \Delta_{\max}}{2}.$$

Note that the above observation is useful when $\frac{\Delta_{\max}}{\Delta_{\min}}$ is bounded above by a constant. So, in Lemma B.12, we consider $S \subset H = \{0,1\}^n$ to be such that the pairwise Hamming distance between any two elements in S is at least $\frac{n}{2}$, to have $\frac{\Delta_{\max}}{\Delta_{\min}} \leq 2$ in our context. It is well known that, there exists such a S with $|S| = \Omega(n)$.

Proof of Lemma B.12. We will show that if there exists an algorithm $\mathcal{A}$ that decides $EMD(p,q) \leq \beta_1 n$ or $EMD(p,q) \geq \beta_2 n$ by using t samples from q, then there exists an algorithm $\mathcal{P}$ that decides whether $\|p-q\|_1 \leq \alpha_1$ or $\|p-q\|_1 \geq \alpha_2$ by using t samples from q, where $\alpha_1 = 2\beta_1$ and $\alpha_2 = 4\beta_2$. Note that we have $0 < \beta_1 < \beta_2 < \frac{1-\epsilon_{\ell_1}}{4}$. So, $0 < \alpha_1 < \alpha_2 < 1 - \epsilon_{\ell_1}$, which satisfies the requirement of Theorem B.14.

Algorithm $\mathcal{P}$:

(1) First run algorithm $\mathcal{A}$.

(2) If the output of algorithm $\mathcal{A}$ is $EMD(p,q) \leq \beta_1 n$, algorithm $\mathcal{P}$ returns $\|p-q\|_1 \leq \alpha_1$.

(3) If the output of algorithm $\mathcal{A}$ is $EMD(p,q) \geq \beta_2 n$, algorithm $\mathcal{P}$ returns $\|p-q\|_1 \geq \alpha_2$.

To complete the proof, we only need to show that $\mathcal{P}$ gives desired output with probability at least $2/3$. The result then follows from Theorem B.14.

Let us first consider the case $\|p-q\|_1 \leq \alpha_1$. Then by Observation B.15, we can say that $EMD(p,q) \leq \frac{\alpha_1 n}{2} = \beta_1 n$. Therefore algorithm $\mathcal{A}$ will output that $EMD(p,q) \leq \beta_1 n$. This implies that the algorithm $\mathcal{P}$ will output $\|p-q\|_1 \leq \alpha_1$.

Now, let us consider the case $\|p-q\|_1 \geq \alpha_2$. Using the fact that any pair elements in $S \subset H$ is at least $\frac{n}{2}$ along with Observation B.15, we get $EMD(p,q) \geq \frac{\alpha_2 n}{4} = \beta_2 n$. This implies $\mathcal{P}$ will output $\|p-q\|_1 \geq \alpha_2$. ☐

Till now, we were discussing the proof of Lemma B.12 that states $\mathrm{QWR}_{\mathrm{EMD}}(n) = \Omega(\frac{n}{\log n})$. The lower bound is almost tight, up to a polynomial factor of $\log n$. The upper bound is stated in the following observation.

Observation B.16. $\mathrm{QWR}_{\mathrm{EMD}}(n) = \widetilde{\mathcal{O}}(n)$, where $\widetilde{\mathcal{O}}(\cdot)$ hides a polynomial factor in $\frac{1}{\beta_2 - \beta_1}$ and $\log n$.

Instead of proving the above observation, we prove the following lemma that states the upper bound of tolerant EMD testing between two distributions when we know one distribution and have sample access to the unknown distribution. By Remark 6, we will be done with the proof of Observation B.16.

Lemma B.17. *Let $H = \{0,1\}^n$ be a n-dimensional Hamming cube, and let p and q denote two known and unknown n-grained distribution over H. There exists an algorithm that takes two parameters β_1, β_2 with $0 \leq \beta_1 < \beta_2 \leq 1$ and a $\delta \in (0,1)$ as input and decides whether $EMD(p,q) \leq \beta_1 n$ or $EMD(p,q) \geq \beta_2 n$ with probability at least $1-\delta$. Moreover, the algorithm* ALG-EMD *queries for $\widetilde{\mathcal{O}}(n)$ many samples from q, where $\widetilde{\mathcal{O}}(\cdot)$ hides a polynomial factor in $\frac{1}{\beta_2-\beta_1}$ and $\log n$.*

Proof. Let ϵ be a constant less than $(\beta_2 - \beta_1)$. We construct a probability distribution q' such that the ℓ_1 distance between q and q' will be at most ϵ, that is, $\sum\limits_{i \in [L]} |q(i) - q'(i)| \leq \epsilon$. Note that such a q' can be constructed with probability at least $1-\delta$ by querying for $\widetilde{\mathcal{O}}\left(n\right)$ many samples of q which follows from [DL12]. Then, we find $EMD(p,q')$. Observe that $|EMD(p,q) - EMD(p,q')| \leq \frac{\epsilon n}{2}$. This is because

$$\begin{aligned}
\left|EMD(p,q) - EMD(p,q')\right| &\leq \left|EMD(p,q') + EMD(q',q) - EMD(p,q')\right| \\
&\leq EMD(q,q') \\
&\leq \frac{\epsilon d}{2} \text{ (By Proposition B.15)}
\end{aligned}$$

As $EMD(p,q) \leq \beta_1 n$ or $EMD(p,q) \geq \beta_2 n$, by the above observation, we will get either $EMD(p,q') \leq \left(\beta_1 + \frac{\epsilon}{2}\right) n$ or $EMD(p,q') \geq \left(\beta_1 + \frac{\epsilon}{2}\right) n$, respectively. By our choice of $\epsilon < \beta_2 - \beta_1$, we can decide $EMD(p,q) \leq \beta_1 n$ or $EMD(p,q) \geq \beta_2 n$ from the value of $EMD(p,q')$.

$\square$

C Missing proofs of Section 3

C.1 Proof of Lemma 3.3

Lemma C.1 (Lemma 3.3 restated)**.** *Let $\kappa \in (0,1)$ and $s \geq 3$ be given constants. Then for $C_{\kappa,s} = \lceil \frac{6s}{\kappa(2-\kappa)} \rceil$ and* sufficiently *large $n \in \mathbb{N}$* [14]*, there exists a graph G_p with $C_{\kappa,s} n$ many vertices such that the following conditions hold.*

(i) *The degree of each vertex in G_p is at least $((1-\kappa)C_{\kappa,s} + 1)n - 1$.*

(ii) *The cardinality of symmetric difference between the sets of neighbors of any two (distinct) vertices in G_p is at least $sn - 2$.*

Proof. To prove the claim, we use probabilistic method to show the existence of a graph G'_p, with $V(G'_p) = C_{\kappa,s} n$, that can have (possible) self loops and satisfy the followings.

(i) The degree of each vertex in G'_p is at least $((1-\kappa)C_{\kappa,s} + 1)n$.

(ii) The cardinality of symmetric difference between the sets of neighbors of any two (distinct) vertices in G'_p is at least sn.

Let us construct a random graph having the vertex set $V(G'_p)$ such that each pair $\{u,v\}$, with $u, v \in V(G'_p)$, is an edge with probability $1 - \frac{\kappa}{2}$ independent of other pairs.

[14] The lower bound of n is a constant that depends on κ and s.

Now we compute the probability that the degree of a vertex $v \in G(V'_p)$, that is $\deg_{G'_p}(v)$, is at most $((1-\kappa)\, C_{\kappa,s}+1)\, n$. For each $v' \in V(G'_p)$, let $X_{v'}$ be the indicator random variable that takes value 1 if and only if $\{v, v'\} \in E(G'_p)$. Note that $\deg_{G'_p}(v) = \sum_{v' \in V(G'_p)} X_{v'}$. Also, $\mathbb{P}(X_{v'}=1) = 1-\frac{\kappa}{2}$. So, the expected value of $\deg_{G'_p}(v)$ is $\left(1-\frac{\kappa}{2}\right) C_{\kappa,s} n$. By using Chernoff bound E.1, we have

$$\begin{aligned}
&\mathbb{P}\left(\deg_{G'_p}(v) \le ((1-\kappa)\, C_{\kappa,s}+1)\, n\right) \\
&= \mathbb{P}\left(\deg_{G'_p}(v) \le (1-\epsilon)\left(1-\frac{\kappa}{2}\right) C_{\kappa,s} n\right) \quad \left(\text{where } \epsilon = \frac{\kappa C_{\kappa,s}-2}{(2-\kappa)C_{\kappa,s}} < 1\right) \\
&\le e^{-\frac{\epsilon^2(2-\kappa)C_{\kappa,s}n}{6}}
\end{aligned}$$

Let $\mathcal{E}_1$ be the event that there exists a vertex $v \in V(G'_p)$ such that the degree of v in G'_p is at most $((1-\kappa)C_{\kappa,s}+1)n$. Using union bound, we can say that $\mathbb{P}(\mathcal{E}_1) \le \left|V(G'_p)\right| e^{-\frac{\epsilon^2(2-\kappa)C_{\kappa,s}n}{6}} \le C_{\kappa,s} n \cdot e^{-\frac{\epsilon^2(2-\kappa)C_{\kappa,s}n}{6}}$. Let $\mathcal{E}_2$ be the event that there exists two (distinct) vertices u, v with $\left|N_{G'_p}(u) \Delta N_{G'_p}(v)\right| \ge sn$, where $N_{G'_p}(u)$ denotes the set of neighbors of u in G'_p. Our goal is to show that G'_p exists which satisfies the required conditions. Observe that, G'_p satisfies the required conditions if and only if $\mathbb{P}(\mathcal{E}_1^c \cap \mathcal{E}_2^c) > 0$. The rest of the work in this proof is to show $\mathbb{P}(\mathcal{E}_1^c \cap \mathcal{E}_2^c) > 0$.

To bound $\mathbb{P}(\mathcal{E}_2)$, consider two distinct vertices u and v. For $w \in V(G'_p)$, let Y_w be the indicator random variable that takes value 1 if and only if $w \in N_{G'_p}(u) \Delta N_{G'_p}(v)$. Note that $\left|N_{G'_p}(u) \Delta N_{G'_p}(v)\right| = \sum_{w \in V(G'_p)} Y_w$ and $\mathbb{P}(Y_w = 1) = 2 \cdot \frac{\kappa}{2}\left(1-\frac{\kappa}{2}\right)$. So, the expected value of $\left|N_{G'_p}(u) \Delta N_{G'_p}(v)\right|$, that is,

$$\mathbb{E}\left[\left|N_{G'_p}(u) \Delta N_{G'_p}(v)\right|\right] = 2 \cdot \frac{\kappa}{2}\left(1-\frac{\kappa}{2}\right) C_{\kappa,s} n.$$

As $C_{\kappa,s} = \lceil \frac{6s}{\kappa(2-\kappa)} \rceil$, $\mathbb{E}\left[\left|N_{G'_p}(u)\ \Delta\ N_{G'_p}(v)\right|\right] \ge 3sn$. Using Chernoff bound E.1, we have

$$\mathbb{P}\left(\left|N_{G'_p}(u)\ \Delta\ N_{G'_p}(v)\right| \le sn\right) \le e^{-\frac{4sn}{9}}$$

Now, by using union bound, we can say that $\mathbb{P}(\mathcal{E}_2) \le \left|V(G'_p)\right|^2 e^{-\frac{4sn}{9}} = C_{\kappa,s}^2 n^2 e^{-\frac{4sn}{9}}$. Finally using union bound one more time and the fact that n is sufficiently large, we have

$$\mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) \le C_{\kappa,s} n \cdot e^{-\frac{\epsilon^2(2-\kappa)C_{\kappa,s}n}{6}} + C_{\kappa,s}^2 n^2 e^{-\frac{4sn}{9}} < 1.$$

Hence, $\mathbb{P}(\mathcal{E}_1^c \cap \mathcal{E}_2^c) > 0$.

□

C.2 Proof of Inequality 2 of Lemma 3.6

Here we prove that

$$d_\phi(G_k, G_u) \le d_\psi(G_k, G_u) + 4x\,|A_k| + 4x + 2y\,|A_k| - y(3n-2). \tag{13}$$

To obtain Inequality (13), let us first consider the case when $x = 1$ and $y = 0$. So, let us assume that $a_i \in A_k$, $a'_j \in A_u$, $b_s \in B_k$ and $b'_s \in B_u$ be such that the following holds: $\psi(a_i) = b'_s$ and $\psi(b_s) = a'_j$, $\psi(x) \in A_u$ for each $x \in A_k \setminus \{a_i\}$, and $\phi(b_t) = b'_t \in B_u$ for each $b_t \in B_k \setminus \{b_s\}$. By the description of Steps (i), (ii) and (iii) of generating ϕ from ψ, as discussed in Lemma 3.6, we have the following observation.

Observation C.2. For $x = 1$ and $y = 0$, we have $\psi(a_i) = b'_s$ and $\psi(b_s) = a'_j$; $\phi(a_i) = a'_j$ and $\phi(b_s) = b'_s$; For any $x \in (A_k \cup B_k) \setminus \{a_i, b_s\}$, $\phi(x) = \psi(x)$.

We can think of ϕ is generated by performing a *swap* operation, that means, the mappings of a_i and b_s are swapped while generating ϕ from ψ. Now we show (for the special case of $x = 1$ and $y = 0$) that:

$$d_\phi(G_k, G_u) \leq d_\psi(G_k, G_u) + 4(|A_k| + 1). \tag{14}$$

By Observation C.2, $\phi(x) = \psi(x)$ for all vertices $x \in (A_k \cup B_k) \setminus \{a_i, b_s\}$. So, any pair of vertices in $(A_k \cup B_k) \setminus \{a_i, b_s\}$ has no effect on $d_\phi(G_u, G_k) - d_\psi(G_u, G_k)$. Following Definition 1.1 and Definition A.2, we can say that

$$d_\phi(G_u, G_k) - d_\psi(G_u, G_k) \leq 2\left[\left|\textsc{Decider}_\phi(a_i)\right| - \left|\textsc{Decider}_\psi(a_i)\right| + \left|\textsc{Decider}_\phi(b_s)\right| - \left|\textsc{Decider}_\psi(b_s)\right|\right]$$

Note that the first term above can be written as $\textsc{Decider}_\phi(a_i) = (\textsc{Decider}_\phi(a_i) \cap (A_k \cup \{b_s\})) \cup (\textsc{Decider}_\phi(a_i) \cap (B_k \setminus \{b_s\}))$. Breaking other terms in the above expression similarly, we have

$$\begin{aligned}
& d_\phi(G_u, G_k) - d_\psi(G_u, G_k) \\
\leq\ & 2[2\,(|A_k| + 1) + \left|\textsc{Decider}_\phi(a_i) \cap (B_k \setminus \{b_s\})\right| - \left|\textsc{Decider}_\psi(a_i) \cap (B_k \setminus \{b_s\})\right| \\
& \quad + \left|\textsc{Decider}_\phi(b_s) \cap (B_k \setminus \{b_s\})\right| - \left|\textsc{Decider}_\psi(b_s) \cap (B_k \setminus \{b_s\})\right|] \\
=\ & 4\,|A_k| + 4 + 2Z, \text{where} \\
& \quad Z = \left|\textsc{Decider}_\phi(a_i) \cap (B_k \setminus \{b_s\})\right| - \left|\textsc{Decider}_\psi(a_i) \cap (B_k \setminus \{b_s\})\right| \\
& \quad\quad + \left|\textsc{Decider}_\phi(b_s) \cap (B_k \setminus \{b_s\})\right| - \left|\textsc{Decider}_\psi(b_s) \cap (B_k \setminus \{b_s\})\right|
\end{aligned}$$

By showing $Z \leq 0$, we will be done with the proof of Inequality (14). Observe that we can say $\textsc{Decider}_\phi(a_i) \cap (B_k \setminus \{b_s\}) = \phi\left(N_{B_k}(a_i)\right) \Delta N_{B_u \setminus \{b'_s\}}(\phi(a_i))$. Also, writing the other terms in the expression of Z in the similar fashion, we get

$$\begin{aligned}
Z \quad \leq \quad & \left|\phi(N_{B_k \setminus \{b_s\}}(a_i)) \Delta \left(N_{B_u \setminus \{b'_s\}}(\phi(a_i))\right)\right| - \left|\psi\left(N_{B_k \setminus \{b_s\}}(a_i)\right) \Delta \left(N_{B_u \setminus \{b'_s\}}(\psi(a_i))\right)\right| \\
& + \left|\phi\left(N_{B_k \setminus \{b_s\}}(b_s)\right) \Delta \left(N_{B_u \setminus \{b'_s\}}(\phi(b_s))\right)\right| - \left|\psi\left(N_{B_k \setminus \{b_s\}}(b_s)\right) \Delta \left(N_{B_u \setminus \{b'_s\}}(\psi(b_s))\right)\right|
\end{aligned}$$

Once again, from Observation C.2,

$$\begin{aligned}
\phi(N_{B_k \setminus \{b_s\}}(a_i)) &= \psi(N_{B_k \setminus \{b_s\}}(a_i)) \text{ (Say } I_1) \\
N_{B_u \setminus \{b'_s\}}(\phi(a_i)) &= N_{B_u \setminus \{b'_s\}}(\psi(b_s)) \text{ (Say } I_2) \\
\phi(N_{B_k \setminus \{b_s\}}(b_s)) &= \phi\left(N_{B_k \setminus \{b_s\}}(b_s)\right) \text{ (Say } I_3) \\
N_{B_u \setminus \{b'_s\}}(\psi(a_i)) &= N_{B_u \setminus \{b'_s\}}(\phi(b_s)) \text{ (Say } I_4)
\end{aligned}$$

Hence, the upper bound on Z can be expressed as follows:

$$\begin{aligned}
Z \ &\leq\ |I_1 \Delta I_2| - |I_1 \Delta I_4| + |I_3 \Delta I_4| - |I_3 \Delta I_2| \\
&\leq\ (|I_1 \Delta I_4| + |I_2 \Delta I_4|) - |I_1 \Delta I_4| + |I_3 \Delta I_4| - |I_3 \Delta I_2| \\
&\leq\ 0
\end{aligned}$$

Here the first two inequalities follow from the triangle inequality.

Note that we were discussing the proof of Inequality (14), which is a special case of Inequality (13) when $x = 1$ and $y = 0$. Observe that, the proof of Inequality (14) does not use any structure of the subgraphs induced by A_k and B_k that changes while performing the swap operation. To prove Inequality (13), we can think of generating ϕ from ψ, by first performing x many swap operations, to generate an intermediate bijection ϕ_1 such that

$$d_{\phi_1}(G_k, G_u) \leq d_\psi(G_k, G_u) + 4x(|A_k| + 1).$$

Observe that $\phi_1(A_k) = A_u$ and $\phi_1(B_k) = B_u$. Then we generate ϕ from ϕ_1, such that ϕ is a SPECIAL bijection, that is, $\phi(b_i) = b'_i$ for each $b_i \in B_k$ along with $\phi_1(A_k) = A_u$ and $\phi_1(B_k) = B_u$ as follows. The process of generation of ϕ from ϕ_1, can be thought of, as if, we are performing y many swap operation between mappings of the vertices in B_{BN}. The difference between, the distance between G_u and G_k w.r.t. the bijections after and before each of the above swaps, is at most $2|A_k| - (3n-2)$. The term $3n-2$ comes from the structure of $G[B_k]$ and $G[B_u]$. Since $|B_{BN}| = y$, $d_{\phi_1}(G_k, G_u) - d_\phi(G_k, G_u)$ is at most $2y|A_k| - y(3n-2)$. Also, we have argued that $d_{\phi_1}(G_k, G_u) \leq d_\psi(G_k, G_u) + 4x(|A_k| + 1)$. Hence, we can finally say that

$$d_\phi(G_k, G_u) \leq d_\psi(G_k, G_u) + 4x(|A_k| + 1) + 2y\,|A_k| - y(3n-2).$$

D Missing Proofs of Section 4

D.1 Proof of Observation 4.12

Observation D.1 (Observation 4.12 restated). If $\left|\mathrm{Symm}_{\phi\phi'}(x)\right| \geq \frac{\gamma_2-\gamma_1}{1000}n$, then

$$\mathbb{P}\left(\left|\mathrm{Symm}_{\phi\phi'}(x) \cap C_u\right| \geq (1 - \frac{1}{50})\left|\mathrm{Symm}_{\phi\phi'}(x)\right| \frac{|C_u|}{n}\right) \leq e^{-\mathcal{O}(|C_u|)}.$$

Proof. Since C_u is taken uniformly at random, we can say that

$$\mathbb{E}\left[\left|(\textsc{Decider}_{\phi'}(x)\Delta\textsc{Decider}_\phi(x)) \cap C_u\right|\right] = \left|\textsc{Decider}_{\phi'}(x)\Delta\textsc{Decider}_\phi(x)\right| \frac{|C_u|}{n}$$

So, using the Chernoff bound mentioned in Lemma E.1, we can say that

$$\mathbb{P}\left(\left|(\textsc{Decider}_{\phi'}(x)\Delta\textsc{Decider}_\phi(x)) \cap C_u\right| \geq \frac{49}{50}\left|\textsc{Decider}_{\phi'}(x)\Delta\textsc{Decider}_\phi(x)\right| \frac{|C_u|}{n}\right)$$
$$\leq e^{-\mathcal{O}(|C_u|)}$$

□

D.2 Proof of Observation 4.13

Observation D.2 (Observation 4.13 restated). (i) If $\left|\textsc{Decider}_\phi(x)\right| \geq \frac{\gamma_2-\gamma_1}{1000}n$, then

$$\mathbb{P}\left(\left|\textsc{Decider}_\phi(x) \cap C_u\right| \geq (1 + \frac{1}{50})\left|(\textsc{Decider}_\phi(x)\right| \frac{|C_u|}{n}\right) \leq e^{-\mathcal{O}(|C_u|)}.$$

(ii) If $\left|\textsc{Decider}_{\phi}(x)\right| < \frac{\gamma_2-\gamma_1}{1000} n$, then $\mathbb{P}\left(\left|\textsc{Decider}_{\phi}(x) \cap C_u\right| \geq \frac{\gamma_2-\gamma_1}{750}\left|C_u\right|\right) \leq e^{-\mathcal{O}(|C_u|)}$.

Proof. (i) Since C_u is taken uniformly at random, we have

$$\mathbb{E}\left[\left|(\textsc{Decider}_{\phi}(x) \cap C_u\right|\right] = \left|\textsc{Decider}_{\phi}(x)\right| \frac{|C_u|}{n}.$$

So, using the Chernoff bound mentioned in Lemma E.1, we have

$$\mathbb{P}\left(\left|\textsc{Decider}_{\phi}(x)\right| \geq \frac{51}{50}\left|\textsc{Decider}_{\phi}(x)\right| \frac{|C_u|}{n}\right) \quad \leq \quad e^{-\mathcal{O}(|C_u|)}$$

(ii) Since C_u is taken uniformly at random, we have

$$\mathbb{E}\left[\left|(\textsc{Decider}_{\phi}(x) \cap C_u\right|\right] \leq \left(\frac{\gamma_2-\gamma_1}{1000}\right)\left|C_u\right|.$$

So, using the Chernoff bound mentioned in Lemma E.1, we have

$$\mathbb{P}\left(\left|\textsc{Decider}_{\phi}(x) \cap C_u\right| \geq \left(\frac{\gamma_2-\gamma_1}{750}\right)\left|C_u\right|\right) \quad \leq \quad e^{-\mathcal{O}(|C_u|)}$$

□

E Some probability results

Lemma E.1 (Chernoff-Hoeffding bound, see [DP09])**.** *Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \in [0,1]$. For $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X]$, the following holds for all $0 \leq \delta \leq 1$*

$$\mathbb{P}\left(|X-\mu| \geq \delta\mu\right) \leq 2\exp\left(\frac{-\mu\delta^2}{3}\right).$$

Lemma E.2 (Chernoff-Hoeffding bound, see [DP09])**.** *Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \in [0,1]$. For $X = \sum_{i=1}^{n} X_i$ and $\mu_l \leq \mathbb{E}[X] \leq \mu_h$, the followings hold for any $\delta > 0$.*

(i) $\mathbb{P}\left(X \geq \mu_h + \delta\right) \leq \exp\left(\frac{-2\delta^2}{n}\right)$.

(ii) $\mathbb{P}\left(X \leq \mu_l - \delta\right) \leq \exp\left(\frac{-2\delta^2}{n}\right)$.

Lemma E.3 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ and $X = \sum_{i=1}^{n} X_i$. Then, for all $\delta > 0$,*

$$\mathbb{P}\left(|X-\mathbb{E}[X]| \geq \delta\right) \leq 2\exp\left(\frac{-2\delta^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$$

.

Lemma E.4 (Theorem 3.2 in [DP09])**.** *Let $X_1, \ldots, X_n$ be random variables such that $a_i \leq X_i \leq b_i$ and $X = \sum_{i=1}^{n} X_i$. Let $\mathcal{D}$ be the* dependent *graph, with vertex set $V(\mathcal{D}) = \{X_1, \ldots, X_n\}$ and edge set $E(\mathcal{D}) = \{(X_i, X_j) : X_i \text{ and } X_j \text{ are dependent}\}$. Then, for all $\delta > 0$,*

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq \delta\right) \leq 2 \exp\left(\frac{-2\delta^2}{\chi^*(\mathcal{D}) \sum_{i=1}^{n} (b_i - a_i)^2}\right),$$

where $\chi^*(\mathcal{D})$ denotes the fractional chromatic number *of $\mathcal{D}$.*

The following lemma directly follows from Lemma E.4.

Lemma E.5 (Chernoff bound for bounded dependency)**.** *Let $X_1, \ldots, X_n$ be indicator random variables such that there are at most d many X_j's on which an X_i depends . For $X = \sum_{i=1}^{n} X_i$ and $\mu_l \leq \mathbb{E}[X] \leq \mu_h$, the followings hold for any $\delta > 0$.*

(i) $\mathbb{P}\left(X \geq \mu_h + \delta\right) \leq \exp\left(\frac{-2\delta^2}{(d+1)n}\right)$,

(ii) $\mathbb{P}\left(X \leq \mu_\ell - \delta\right) \leq \exp\left(\frac{-2\delta^2}{(d+1)n}\right)$.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il