Interplay between Graph Isomorphism and Earth Mover’s Distance in the Query and Communication Worlds

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 \(\gamma_1\) or more than \(\gamma_2\), where \(\gamma_1\) and \(\gamma_2\) 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 \(\gamma_1\) is 0) has been studied by Fischer and Matsliah (SICOMP’08).

In this paper, we prove a (interesting) connection between tolerant graph isomorphism testing and tolerant testing of the well studied Earth Mover’s Distance (EMD). We prove that deciding tolerant graph isomorphism is equivalent to deciding tolerant EMD testing between multi-sets in the query setting. Moreover, the reductions between tolerant graph isomorphism and tolerant EMD testing (in query setting) can also be extended directly to work in the two party Alice-Bob communication model (where Alice and Bob have one graph each and they want to solve tolerant graph isomorphism problem by communicating bits), and possibly in other sublinear models as well.

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) 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.

*A preliminary version of this paper has appeared in RANDOM 2021.
†Indian Statistical Institute, Kolkata, India
‡University of Warwick, UK. This work was done when the author was a graduate student at Indian Statistical Institute, Kolkata, India.
## Contents

1 Introduction ................................. 1
   1.1 Query and Communication Models for Graph Isomorphism .................. 1
      1.1.1 Query Complexity of Graph Isomorphism .......................... 2
      1.1.2 Communication Complexity of Graph Isomorphism .................. 3
   1.2 Earth Mover’s Distance (EMD) ........................................... 3
   1.3 Our Results ............................................................... 5

2 Preliminaries ................................. 6
   2.1 Notion of distance between two graphs ................................... 6
   2.2 Property Testing of Distribution Properties ............................. 7
   2.3 Earth Mover’s Distance (EMD) over Hamming Cube ........................ 7
   2.4 Communication Complexity Landscape of GI ................................ 12
   2.5 Useful concentration bounds ............................................. 13

3 Discussion on our proof of Theorem 1.4 ....................................... 13
   3.1 Lower bound part of Theorem 1.4 ....................................... 13
   3.2 Upper bound part of Theorem 1.4 ....................................... 15

4 Tolerant GI is as hard as tolerant EMD testing ................................. 16
   4.1 Tolerant GI to Tolerant EMD testing: Proof of Lemma 4.2 ................. 17

5 Tolerant EMD testing is as hard as tolerant GI ................................ 24
   5.1 Algorithm for tolerant GI testing ........................................ 24
      5.1.1 Phase 1 ..................................................................... 25
      5.1.2 Phase 2 ..................................................................... 25
      5.1.3 Phase 3 ..................................................................... 26
   5.2 Proof of correctness ........................................................... 27
      5.2.1 Proof of Completeness Property ...................................... 27
      5.2.2 Proof of Soundness Property ......................................... 34
   5.3 Proof of Theorem 5.1 ........................................................... 37

6 Communication Complexity Landscape of GI .................................... 37
   6.1 Deterministic Communication Complexity of GI ........................... 37
   6.2 Randomized Communication Complexity of Non-tolerant GI ............. 39

7 Conclusion ........................................... 40
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 $\psi : V(G_u) \rightarrow V(G_k)$ 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 $\{\psi(u), \psi(v)\} \in E(G_k)$. 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 $\phi : V_u \rightarrow V_k$, the distance between the graphs $G_u$ and $G_k$ with respect to the bijection $\phi$ 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 $\min_{\phi} d_\phi(G_u, G_k)/n^2$, 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 (see Section 1.1.1) and two party communication complexity model (see Section 1.1.2). In this work, our main contribution is to show a unified connection between graph isomorphism testing and Earth Mover’s distance testing which holds across computational models. We present our results in Section 1.3 after we introduce the Earth Mover’s distance in Section 1.2.

1.1 Query and Communication Models for 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 $\alpha$ such that

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

In a graph $G$, $V(G)$ and $E(G)$ denote the sets of vertices and edges in $G$, respectively.
We now present the two models (query and communication) for studying the above problem.

1.1.1 Query Complexity of Graph Isomorphism

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 $0 < \gamma < 1$, two graphs $G_u$ and $G_k$, with $n$ vertices, are called $\gamma$-close or $\gamma$-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 $(\gamma_1, \gamma_2)$-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 $\gamma_1$-close or $\gamma_2$-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 $G_k$) 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 $2/3$\(^3\)), whether $G_u$ and $G_k$ are $\gamma_1$-close or $\gamma_2$-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. Recently, the non-tolerant testing of GI has been considered in various other models (like Goldreich [Gol19] studied the problem for the bounded degree graph model of property testing and Levi and Medina [LM20] considered the problem in the distributed setting). However, the tolerant version of the problem remains elusive and it is surprising that the tolerant version of a fundamental problem like graph isomorphism (in query model) is not addressed in the literature, though the non-tolerant version of GI testing problem has been resolved more than a decade ago in [FM08] (when one graph is unknown). On a different note, there are also studies of non-tolerant version of graph isomorphism testing in the literature when both the graphs are unknown [FM08, OS18]. We will not discuss much about that case as the main focus of this paper is different.

Before proceeding further, we want to note that there is a simple algorithm with query complexity $\tilde{O}(n)$ for tolerant testing of graph isomorphism (when one of the graphs is known in advance). Basically, one goes over all possible $n!$ bijections $\phi : V_u \rightarrow V_k$ and estimates the distance between $G_u$ and $G_k$ with respect to the permutation. The samples may be reused\(^5\) and hence we have the following observation.

**Observation 1.2.** 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 $O(n)$ queries and outputs a number $\alpha$ such that, with probability at least $\frac{2}{3}$, the following holds:

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

But obtaining a lower bound matching (at least up to a polylog factor) the upper bound of Observation 1.2 is not at all obvious. This paper’s main contribution is to show an equivalence between tolerant testing of graph isomorphism and tolerant EMD testing between multi-sets (in the query setting).

Like many other property testing problems, the core difficulty in the testing of GI is understanding certain properties of distributions. In the case of the non-tolerant version of GI, it has been shown in

\(^2\)As a shorthand, rather than saying $\gamma$-close or $\gamma$-far to isomorphic, we will just say $\gamma$-close or $\gamma$-far respectively.

\(^3\)Specifically, $G_u$ 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.

\(^5\)If the samples are $\Theta(\log(n!))$, then the error probability can be bounded using the union bound.
that the core problem is testing the variation distance between two distributions. 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 \( n \) \(^6\) then GI can be tested using \( \tilde{O}(q(n)) \) queries, where the tilde hides a polylogarithmic factor of \( n \) (number of vertices). And since the query complexity for testing identity of distributions (from \([BFF+01]\), \([Pan08]\), \([ADK15]\), \([VV17]\)) is known to be \( O(\sqrt{n}\epsilon^2) \), the query complexity for non tolerant GI-testing is \( \tilde{O}(\sqrt{n}) \).

In the lower bound proof of \([FM08]\), there is 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 testing variation distance between two distributions.

1.1.2 Communication Complexity of Graph Isomorphism

One of the central models of computation (particularly in the context of theoretical computer science) is the 2-player communication game introduced by Yao \([Yao79]\). Communication complexity is one of the most studied complexity measures and has wide-ranging applications in many different areas of computer science. But surprisingly, as far as we know, the communication complexity problem of GI (where Alice has graph \( G_a \) and Bob has graph \( G_b \), and they want to decide if \( G_a \) and \( G_b \) are isomorphic) has never been studied. One of the main reasons may be that, in the communication setup, the standard GI problem reduces to the string equality checking problem, and hence GI in the (randomized) communication setup is not that interesting anymore, since the randomized communication complexity, trivially, becomes \( O(1) \).

But when it comes to tolerant GI testing, the communication version is not at all obvious. So, if Alice and Bob are given two graphs \( G_a \) and \( G_b \) respectively, what is the (randomized) communication complexity for checking if \( d(G_a, G_b) \leq \gamma_1 n^2 \) or \( d(G_a, G_b) \geq \gamma_2 n^2 \)? We discuss our result on this in Section 6 in detail.

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:

\[
\text{Minimize } \sum_{i,j \in H} f_{ij}d_H(i, j) \quad \text{Subject to } \sum_{j \in H} f_{ij} = p(i) \forall i \in H, \text{ and } \sum_{i \in H} f_{ij} = q(j) \forall j \in H.
\]

A standard way to think of sampling from any probability distribution is to consider it as a 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 1.3 (EMD over multi-sets while sampling with and without replacement).** Let \( S_1 \) and \( S_2 \) denote two multi-sets, over \( n \)-dimensional Hamming cube \( H = \{0, 1\}^n \) such that \( |S_1| = |S_2| = n \). Consider the two distributions \( p_1 \) and \( p_2 \) over the Hamming cube \( H \) that are naturally defined by the sets \( S_1 \) and \( S_2 \) where for all \( x \in H \) probability of \( x \) in \( p_1 \) (and \( p_2 \)) is the

\(^6\)Testing identity between two distributions means to test if the unknown distribution (from where the samples are drawn) is identical to the known distribution or if the variation distance between them more than \( \epsilon \).
number of occurrences of $x$ in $S_1$ (and $S_2$) divided by $n$. We then define the EMD between the multi-sets $S_1$ and $S_2$ as

$$EMD(S_1, S_1) \triangleq n \cdot EMD(p_1, p_2).$$

The problem of estimating the EMD over multi-sets while sampling with (or without) replacement means designing an algorithm, that given any two constants $\beta_1, \beta_2$ such that $0 \leq \beta_1 < \beta_2 \leq 1$, a known multi-set $S_k$ and access to the unknown multi-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 QWR$_{ EDM}(n, \beta_1, \beta_2)$ (and QWoR$_{ EDM}(n, \beta_1, \beta_2)$) 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 QWR$_{ EDM}(n)$ (QWR$_{ EDM}(n)$) instead of QWR$_{ EDM}(n, \beta_1, \beta_2)$ (QWR$_{ EDM}(n, \beta_1, \beta_2)$) 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 $\ell_1$ metric. Even the problem of estimation of EMD between distributions up to an additive factor has been well studied, for reference see [DBNNR11, SPT18]. 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 $\epsilon$) 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 distributions when the metric space is the Hamming cube. For a formal discussion on EMD over the Hamming cube, please refer to Section 2.3.

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 $\Omega(\frac{n}{\log n})$. In other words, QWR$_{ EDM}(n) = \Omega(\frac{n}{\log n})$, and this is tight ignoring polynomial factor in $\log n$ (See Theorem 2.13 of Section 2.3). But what about QWoR$_{ EDM}(n)$? 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 (for testing/estimating distances between distributions/multi-sets). However, it is interesting to note that, sampling without replacement model has been considered before in a different context by Raskhodnikova, Ron, Shpilka and Smith [RRSS09] for proving a lower bound of distinct elements problem. Also, recently Goldreich [Gol19] considered a similar sampling without replacement model while studying the non-tolerant graph isomorphism in the bounded degree model.

Coming back to our context: Observe that if QWoR$_{ EDM}(n) = O(\sqrt{n})$, see Proposition 2.10 in Section 2.3. As QWR$_{ EDM}(n) = \Omega(\frac{n}{\log n})$, we have a lower bound of $\Omega(\sqrt{n})$ on QWoR$_{ EDM}(n)$. To the best of our knowledge, there is no known better lower bound than $\Omega(\sqrt{n})$ for QWoR$_{ EDM}(n)$, although a lower bound of $\Omega(\frac{n}{\log n})$ exists for QWR$_{ EDM}(n)$ (using observation in [DBNNR11]). Note that QWR$_{ EDM}(n) = \tilde{O}(n)$.

Note that the main contribution of our work is the introduction of the complexity measure QWoR$_{ EDM}(n)$ and its connection to graph isomorphism testing across query and communication models.
1.3 Our Results

Our main technical 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 1.4** (Result in Query Complexity). 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 $\gamma_1, \gamma_2$ 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 $\frac{k}{3}$. Then

$$Q_{GI}(G_u, G_k) = \tilde{\Theta}(QW\text{O}R_{\text{EMD}}(n))$$

where $\tilde{\Theta}(\cdot)$ hides polynomial factors in $\frac{1}{\sqrt[72]{\gamma_1}}$ and $\log n$.

This gives us a geometric approach for solving the graph isomorphism testing problem. Thus improving the bound of $QW\text{O}R_{\text{EMD}}(n)$ would directly provide us a better bound on $Q_{GI}(G_u, G_k)$.

On the other hand, extending the lower bound of $QW\text{R}_{\text{EMD}}(n)$ to $Q\text{W}\text{O}R_{\text{EMD}}(n)$ would give us a better lower bound on $Q_{GI}(G_u, G_k)$. However, the difference between sampling with and without replacement is much more subtle. Freedman [Fre77] has shown the difference when we sample elements with replacement from a set and that without replacement from the same set. However, when the number of samples is $o(\sqrt{n})$, the distribution of answers to the queries when samples are drawn with replacement is very close (in $\ell_1$ distance) to the distribution of answers to the queries when samples are drawn without replacement. Thus, following the simulation of samples with replacement using samples without replacement (stated formally in Proposition 2.10) along with Theorem 1.4, 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 $\gamma$-far from isomorphic, with $\gamma \leq \zeta$, must make $\Omega(\sqrt{n})$ queries.

Our proof of Theorem 1.4 has two parts: for the lower bound, we reduce tolerant testing of EMD of multi-sets over the Hamming cube using samples without to tolerant graph isomorphism testing. For the upper bound, we reduce from tolerant graph isomorphism to tolerant testing of EMD of multi-sets over the Hamming cube using samples without replacement. Thus our reductions also hold for other computational models such as the communication complexity model. We state our result in communication setting below.

**Theorem 1.6** (Result in Communication Complexity). If Alice and Bob are given two graphs $G_a$ and $G_b$ with $n$ vertices respectively and the (randomized) communication complexity for checking if the graphs are $\gamma_1$-close or $\gamma_2$-far is $c(n, \gamma_1, \gamma_2)$ then the following holds: There exists an absolute constant $C$ such that if Alice and Bob are given two $n$-grained distributions over the $Cn$-dimension Hamming cube, then the (randomized) communication complexity of checking if the Earth Mover’s Distance between the distributions is at most $\beta_1 n$ or at least $\beta_2 n$ is $\tilde{\Theta}(c(n, \gamma_1', \gamma_2'))$, where $\gamma_1'$ and $\gamma_2'$ are constants that depend only on $\beta_1$ and $\beta_2$, and $\tilde{\Theta}(\cdot)$ hides multiplicative factor of $\poly(\log n)$.

Theorem 1.6 says that the communication complexity of solving tolerant graph isomorphism and tolerant EMD testing are essentially the same, ignoring the polylog factor. Note that in the case

---

7The probability of each element in the sample space is an integer multiple of $\frac{1}{n}$. 

5
of the communication setting, the distinction between with replacement and without replacement is not present. Also, it is important to point out that the lower bounds on tolerant EMD in the sampling model ([Val11] and [VV11]) does not give a lower bound in the communication setting. Though the tolerant graph isomorphism problem has not been addressed at all in the literature of communication complexity, EMD (for different metric spaces) has been considered in communication, streaming, and sketching models [KN06, AIK08, ADBIW09, AKR18]. However, the EMD problem that we have considered in this paper is different from those considered in the literature, and we believe that it will be of independent interest.

We also observe that the deterministic communication complexity of graph isomorphism is $\Omega(n^2)$ even for the non-tolerant setting.

**Theorem 1.7.** Deterministic communication complexity of non-tolerant version of Graph Isomorphism testing (hence the tolerant version) is $\Theta(n^2)$.

We will prove this theorem in Section 6.

**Organization of the paper**

In Section 3 we discuss the proof techniques of our main results. We prove the lower bound part (tolerant graph isomorphism is as hard as tolerant EMD testing) and upper bound part (tolerant EMD testing is as hard as tolerant graph isomorphism) of Theorem 1.4 in Sections 4 and 5 respectively. We present the results on communication complexity in Section 6. We finally conclude in Section 7.

**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 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 $d_H(x, y)$.

2.1 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 1.4.

**Definition 2.1.** (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 $\phi$ 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,

$$\text{DECIDER}_\phi(x) := \{y \in V(G_u) : \text{one of the edges } \{x, y\} \text{ and } \{\phi(x), \phi(y)\} \text{ is not present}\}$$
Definition 2.2. (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 $\phi$ is defined as the sum of the sizes of the deciders of all the vertices in $G_u$, that is,

$$d_\phi(G_u, G_k) := \sum_{x \in V(G_u)} |\text{DECIDER}_\phi(x)|.$$ 

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

Remark 1. 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 $d(G_u, G_k) = 2 \left(\frac{n}{2}\right) \delta_{GI}(G_u, G_k)$. 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 2.3. (Close and Far) For $\gamma \in [0, 1)$, two graphs $G_u$ and $G_k$ with $n$ vertices are $\gamma$-close to isomorphic if $d(G_u, G_k) \leq \gamma n^2$. Otherwise, we say $G_u$ and $G_k$ are $\gamma$-far from being isomorphic.  

2.2 Property Testing of Distribution Properties

Understanding different properties of probability distributions have been an active area of research in property testing (For reference, see Can20). 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.

2.3 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 $\ell_1$ distance between two distributions.

Definition 2.4 ($\ell_1$ distance between two distributions). Let $p$ and $q$ be two probability distributions over $[n]$. The $\ell_1$ distance between $p$ and $q$ is defined as

$$d_{\ell_1}(p, q) = \sum_{i=1}^{n} |p(i) - q(i)|$$

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

8By abuse of notation, we will say $G_u$ and $G_k$ are $\gamma$-far when $d(G_u, G_k) \geq \gamma n^2$.  

7
**Definition 2.5** (*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:

\[
\begin{align*}
\text{Minimize} & \sum_{x, y \in H} f_{xy}d_H(x, y) \\
\text{Subject to} & \sum_{y \in H} f_{xy} = p(x) \forall x \in H, \text{ and } \sum_{x \in H} f_{xy} = q(y) \forall y \in H.
\end{align*}
\]

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

**Definition 2.6** (*$EMD$ 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 $EMD(S_1, S_2) = \min_{\phi: S_1 \rightarrow S_2} \sum_{x \in S_1} d_H(x, \phi(x))$ where $\phi$ is a bijection from $S_1$ to $S_2$.

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

**Definition 2.7** (*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 note that we can simulate samples with replacement when we have samples without replacement.

**Proposition 2.8** (*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 $P(x_i = s_i) = \frac{1}{n}$.

Thus we can simulate $Q$ many samples with replacement from $Q$ many samples without replacement from the unknown multi-set $S_u$. \hfill $\square$

The following observation connects the $EMD$ between two probability distributions with that of between two multi-sets.
Observation 2.9. Let \( p, q \) be two \( K \)-grained probability distributions\(^9\) 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 2.5 and 2.6, 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, \( EMD(p, q) = \sum_{i,j \in H} f_{ij}d_H(i, j) \). Consider a bijection \( \phi \) from \( S_1 \) to \( S_2 \) where \( \phi(i) = j \) for \( g_{ij} \) many \( i \)'s. Hence, by Definition 2.6

\[ EMD(S_1, S_2) \leq \sum_{x \in S_1} d_H(x, \phi(x)) = \sum_{i,j \in H} K \cdot f_{ij}d_H(i, j) = K \cdot EMD(p, q). \]

Now, we show \( K \cdot EMD(p, q) \leq EMD(S_1, S_2) \). Let \( \phi^* \) 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 \( \phi \), 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 2.5

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

Remark 2. Note that sample access from a probability distribution is exactly same as uniform sampling from a multi-set \textbf{with} replacement.

Proposition 2.10. Let \( D \) be the set of all multi-sets of size \( n \) over a universe \([m]\); let \( S_k \) and \( S_u \) in \( D \) denote the known and unknown multi-sets over \([n]\); and \( \text{PROP} : D \times D \rightarrow \{0, 1\} \) be a boolean function. Then the following holds:

If there exists an algorithm that determines \( \text{PROP} \) by \( Q \) many samples \textbf{without} replacement from \( S_u \) with probability at least \( 2/3 \), then there exists an algorithm that determines \( \text{PROP} \) by \( \min\{Q, \sqrt{\min\{n, m\}}\} \) many samples \textbf{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 \([n]\) \textbf{with} \( \textbf{without} \) replacement, the \( \ell_1 \) distance between \( D_{WR} \) and \( D_{WoR} \) is \( o(1) \).

Definition 2.11 (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 \( x \in H \)

\[^9\]The probability of each element in the sample space is an integer multiple of \( \frac{1}{K} \).
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 $\beta_1, \beta_2$ 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 $QWR_{EMD}(n, d, \beta_1, \beta_2)$ (and $QWoR_{EMD}(n, d, \beta_1, \beta_2)$) denote the number of samples with (and without) replacement required to decide the above from the unknown multi-set $S_u$. For ease of presentation, we write $QWoR_{EMD}(n, d)$ ($QWR_{EMD}(n, d)$) instead of $QWoR_{EMD}(n, d, \beta_1, \beta_2)$ when the proximity parameters are clear from the context.

**Proposition 2.12** (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) $QWR_{EMD}(n_1, d) \leq QWR_{EMD}(n_2, d)$;

(ii) $QWoR_{EMD}(n_1, d) \leq QWoR_{EMD}(n_2, d)$;

(iii) $QWR_{EMD}(n, d_1) \leq QWR_{EMD}(n, d_2)$; and

(iv) $QWoR_{EMD}(n, d_1) \leq QWoR_{EMD}(n, d_2)$.

**Remark 3.** For $d = n$ (as considered in Definition 1.3), $QWoR_{EMD}(n, d)$ (and $QWR_{EMD}(n, d)$) are denoted as $QWoR_{EMD}(n)$ (and $QWR_{EMD}(n)$).

Now let us state the lower bound of $QWR_{EMD}(n)$.

**Theorem 2.13.** $QWR_{EMD}(n) = \Omega\left(\frac{n}{\log n}\right)$.

Thus following Proposition 2.10 we have

**Theorem 2.14.** $QWoR_{EMD}(n) = \Omega(\sqrt{n})$.

Note that an upper bound of $QWoR_{EMD}(n) = \tilde{O}(n)$ is trivial. In the rest of the section, we focus on proving Theorem 2.13 that states the lower bound on $QWR_{EMD}(n)$. We also provide an upper bound for $QWR_{EMD}(n)$ at Lemma 2.19 that shows that $\tilde{O}(n)$ many samples with replacement from $S_u$ to estimate $QWR_{EMD}(n)$. Note that by Remark 2 it is enough to show the following lemma that states the lower bound for tolerant EMD testing between two distributions.

**Lemma 2.15.** Let $S$ be a subset of a Hamming cube $H = \{0, 1\}^n$ such that the minimum distance between any pair of points in $S$ is at least $\frac{n}{2}$. Also, let $p$ and $q$ be two known and unknown distributions, respectively, supported over a subset of $S$. Then there exists a constant $\epsilon_{EMD}$ such that the following holds. Given two constants $\beta_1, \beta_2$ with $0 < \beta_1 < \beta_2 < \epsilon_{EMD}$, $\Omega\left(\frac{n}{\log n}\right)$ samples from the distribution $q$ are necessary in order to decide whether $EMD(p, q) \leq \beta_1 n$ or $EMD(p, q) \geq \beta_2 n$. More over, $\epsilon_{EMD} = \frac{1-\epsilon_{\ell_1}}{4}$, where $\epsilon_{\ell_1}$ is the constant that is mentioned in Theorem 2.17.

To prove the above lower bound, let us first consider the following lower bound for tolerant $\ell_1$ testing between two probability distributions.
Theorem 2.16 (Valiant and Valiant [VVII]). Let $p$ and $q$ be two known and unknown probability distributions respectively over $[n]$. There exists an absolute constant $\epsilon$ 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. \footnote{Note that this is rephrasing of the result proved in [VVII]. For reference, see Chapter 5 of the survey by Canonne [Can20].}

Now, we restate the above result for our purpose.

**Theorem 2.17.** Let $p$ and $q$ be two known and unknown probability distributions, having support size $n$, over a Hamming cube $H = \{0,1\}^n$. There exists an absolute constant $\epsilon_\ell_1$ 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 2.13 by using Lemma 2.17. However, Theorem 2.13 is regarding $\text{EMD}$ between two distributions whereas Lemma 2.17 is regarding $\ell_1$ distance between two distributions. The following observation (from [DBNNR11]) gives a connection between $\text{EMD}$ between two distributions with the $\ell_1$ distance between them, which will be required in lower bound proof.

**Proposition 2.18 ([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:

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

Note that the above proposition gives interesting result when $\frac{\Delta_{\max}}{\Delta_{\min}}$ is bounded by a constant. Note that $S \subseteq \{0, 1\}^n$ satisfies $\frac{\Delta_{\max}}{\Delta_{\min}} \leq 2$.

**Proof of Lemma 2.15.** In $S \subseteq H = \{0, 1\}^n$, 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 $|S| = \Omega(n)$. We will show that if there exists an algorithm $A$ that decides $\text{EMD}(p, q) \leq \beta_1 n$ or $\text{EMD}(p, q) \geq \beta_2 n$ by using $t$ samples from $q$, then there exists an algorithm $P$ 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 2.17.

**Algorithm $P$:**

(1) First run algorithm $A$.

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

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

To complete the proof, we only need to show that $P$ gives desired output with probability at least $2/3$. The result then follows from Theorem 2.17.

Let us first consider the case $\|p - q\|_1 \leq \alpha_1$. Then by Observation 2.18, we can say that $\text{EMD}(p, q) \leq \frac{\alpha_1 n}{2} = \beta_1 n$. Therefore algorithm $A$ will output that $\text{EMD}(p, q) \leq \beta_1 n$. This implies that the algorithm $P$ 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 2.18, we get $\text{EMD}(p, q) \geq \frac{\alpha_2 n}{4} = \beta_2 n$. This implies $P$ will output $\|p - q\|_1 \geq \alpha_2$. \qed
Till now, we were discussing the proof of Lemma 2.15 that states \( \text{QWR}_{\text{EMD}}(n) = \Omega(\frac{n}{\log n}) \). The lower bound is almost tight, up to a polynomial factor of \( \log n \). The upper bound is stated in the following observation.

**Observation 2.19.** \( \text{QWR}_{\text{EMD}}(n) = \tilde{O}(n) \), where \( \tilde{O}() \) 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 2 we will be done with the proof of Observation 2.19.

**Lemma 2.20.** 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 \( \beta_1, \beta_2 \) 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 \( \text{ALG-EMD} \) queries for \( \tilde{O}(n) \) many samples from \( q \), where \( \tilde{O}() \) hides a polynomial factor in \( \frac{1}{\beta_2 - \beta_1} \) and \( \log n \).

**Proof.** Let \( \epsilon \) be a constant less than \((\beta_2 - \beta_1)\). We construct a probability distribution \( q' \) such that the \( \ell_1 \) distance between \( q \) and \( q' \) will be at most \( \epsilon \), that is, \( \sum_{i \in [L]} |q(i) - q'(i)| \leq \epsilon \). Note that such a \( q' \) can be constructed with probability at least \( 1 - \delta \) by querying for \( \tilde{O}(n) \) many samples of \( q \) which follows from [DL12]. Then, we find \( EMD(p,q') \). Observe that \( |EMD(p,q) - EMD(p,q')| \leq \frac{\epsilon d}{2} \). This is because

\[
|EMD(p,q) - EMD(p,q')| \leq \left| EMD(p,q') + EMD(q',q) - EMD(p,q') \right| \\
\leq EMD(q,q') \\
\leq \frac{\epsilon d}{2} \quad \text{(By Proposition 2.18)}
\]

As \( EMD(p,q) \leq \beta_1 n \) or \( EMD(p,q) \geq \beta_2 n \), by the above observation, we will get either \( EMD(p,q') \leq (\beta_1 + \frac{\epsilon}{2}) n \) or \( EMD(p,q') \geq (\beta_1 + \frac{\epsilon}{2}) n \), 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 \( EMD(p,q') \).

\( \square \)

### 2.4 Communication Complexity Landscape of GI

Two players Alice and Bob have two graphs \( G_a \) and \( G_b \) (on \( n \) vertices) respectively. They would like to communicate among themselves to decide about the following problems:

1. **Non-tolerant Graph Isomorphism**: If \( G_a \) and \( G_b \) are isomorphic or \( \epsilon \)-far from isomorphic where \( \epsilon \in (0,1) \) is a proximity parameter.

2. **Tolerant Graph Isomorphism**: If \( G_a \) and \( G_b \) are \( \epsilon_1 \)-close to being isomorphic or \( \epsilon_2 \)-far from being isomorphic where \( \epsilon_1, \epsilon_2 \) are two proximity parameters such that \( 0 \leq \epsilon_1 < \epsilon_2 \leq 1 \).

We show that the deterministic communication complexity for both these problems is \( \Theta(n^2) \) in Theorem 6.1. We also prove that the randomized communication complexity for the Non-tolerant graph isomorphism problem is \( O(1) \) (with shared randomness). The communication complexity for the tolerant graph isomorphism remains open. We showed in this paper (Theorem 6.2) that the randomized communication complexity of the Tolerant Graph Isomorphism is same as the randomized communication complexity of the Tolerant EMD problem on the Hamming cube.

The results are summarized in the following table:

The proofs of these results are presented in Section 6.
<table>
<thead>
<tr>
<th>Models</th>
<th>Non-tolerant Graph Isomorphism</th>
<th>Tolerant Graph Isomorphism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deterministic</td>
<td>$\Theta(n^2)$</td>
<td>$\Theta(n^2)$</td>
</tr>
<tr>
<td>Randomized</td>
<td>$\Theta(1)$</td>
<td>OPEN!</td>
</tr>
</tbody>
</table>

Table 1: Summary of GI results in communication complexity.

2.5 Useful concentration bounds

In our work, we use the following three concentration inequalities, see [DP09].

Lemma 2.21 (Chernoff-Hoeffding bound). 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 $0 \leq \delta \leq 1$

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

Lemma 2.22 (Chernoff-Hoeffding bound). 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) $\mathbb{P}(X \geq \mu_h + \delta) \leq \exp \left( \frac{-2\delta^2}{n} \right)$.

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

Lemma 2.23 (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 $\delta > 0$,

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

3 Discussion on our proof of Theorem 1.4

3.1 Lower bound part of Theorem 1.4

In this reduction, we crucially use the fact that the multi-sets are composed of elements from the Hamming cube. The reduction is based upon an involved gadget construction. In fact, we prove the lower bound for a slightly more powerful query model rather than the standard adjacency matrix query model. 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.

Now, we discuss the overview of our reduction. Let $S_k$ and $S_u$ denote the known and the unknown multi-sets, over a Hamming cube $\{0, 1\}^d$ (of dimension $d$) with $d = \Theta(n)$, having $n$ elements each.
To start with, let us assume that we know both $S_k$ and $S_u$. We will construct two graphs $G_k$ and $G_u$ on $d + n$ vertices as follows:

- The vertex set of $G_k$ (and $G_u$) are partitioned into two sets $A_k$ and $B_k$ (and $A_u$ and $B_u$) with $|A_k| = |A_u| = n$ and $|B_k| = |B_u| = d$.

- The graph induced by $A_k$ is a clique, and similarly the graph induced by $A_u$ is a clique.

- The graphs induced by $B_k$ and $B_u$ are copies of a special graph with certain nice properties which enable our reduction to work. The existence of such a graph is proved (in Lemma 4.3) using a probabilistic argument.

- Finally, for the cross edges between $A_k$ and $B_k$ (and $A_u$ and $B_u$), we have: there is an edge between the $i$-th vertex of $A_k$ (or $A_u$) and the $j$-th vertex of $B_k$ (or $B_u$) if and only if the $j$-th coordinate of the $i$-th element of $S_k$ (or $S_u$) is 1.

- Finally, a random permutation $\pi$ is applied to the vertices of $G_u$.

The permutation $\pi$ is not known to the GI-tester. Note that we can construct $G_k$ explicitly as $S_k$ is known. However, that is not the same with $G_u$ as $S_u$ is unknown. But since we know the permutation $\pi$, any query to the adjacency matrix of the graph $G_u$ can be answered by a single query to one bit of $S_u$. But unfortunately we don’t have query access to $S_u$, and only have sample access to $S_u$. To deal with this problem, it is easier to consider a slightly more powerful query. Say, the GI-tester wants to query the $(i,j)$-th bit of the graph $G_u$. Of course, if both $i$ and $j$ are in $A_u$ or both are in $B_u$, we can answer without even sampling from $S_u$. But if $i$ is in $A_u$ and $j$ is in $B_u$, then what we intend to do is to give the whole neighborhood of $i$ in $B_u$ as the answer to the query. This would be like neighbourhood query in a bipartite graph. But the question remains: how do we intend to answer the query by sampling. The key observation here is that since the GI-tester does not know the permutation $\pi$ that was applied to the vertices in $G_u$, to its eye, all the vertices that have not been touched so far look same. So, every time it queries for $(i,j)$, where $i \in A_u$ and $j \in B_u$, either of the two cases can happen:

- Either, previously a query of the form $(i, j_1)$ was asked where $j_1$ is also in $B_u$, but in that case, it must have already got the answer of $(i,j)$ as we must have given all the neighbors of $i$ in $B_u$. So in that case, we can give back the same answer without sampling.

- Or, previously $i$ did not participate in any query of the form $(i, j_1)$ where $j_1$ is in $B_u$. In this case, to the GI-tester’s eye, $i$ is just a new vertex from $A_u$. We can then sample without replacement from $S_u$ and whatever sample of the multi-set we have, we can assume that it is the element $i$ and answer accordingly. Note that this is the exact place where sampling without replacement is crucial.

To complete our proof, we need to prove how the GI-distance between $G_k$ and $G_u$ is connected to the EMD between $S_k$ and $S_u$. Consider the set $\Phi$ of all special bijections from $V(G_k)$ to $V(G_u)$ that maps $A_k$ into $A_u$ and $B_k$ into $B_u$ such that the $i$-th vertex of $B_k$ is mapped to the $i$-th vertex of $B_u$. Observe that $d_{\Phi}(G_k, G_u) = 2 \cdot \text{EMD}(S_k, S_u)$, where $d_{\Phi}(G_k, G_u) = \min_{\phi \in \Phi} d_{\phi}(G_k, G_u)$ (See Lemma 4.5 for a formal proof). The factor 2 is because of the way we define $d_{\phi}(G_k, G_u)$ (See Definition 1.1). This implies that tolerant isomorphism testing between $G_k$ and $G_u$ is at least as hard as tolerant EMD testing between $S_k$ and $S_u$ if we restrict the bijection from $V(G_k)$ to $V(G_u)$ to be a special bijection. The reduction works for all possible bijections, because of the careful choice of the subgraph of $G_k$.
(and \(G_u\)) induced by \(B_k\) (and \(B_u\)), thus ensuring \(d(G_k, G_u)\) is close to \(d_\Phi(G_k, G_u)\) (See Lemma \ref{lemma:approximation} for a formal proof).

One might compare our proof technique to the lower bound proof of (non-tolerant) testing of GI from \cite{FM08}. In \cite{FM08}, \(\Omega(\sqrt{n})\) 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 GI testing to equivalence testing of two probability distributions. But in our lower bound proof, we establish a direct reduction to estimating EMD of multi-sets on the Hamming cube with access to samples \textbf{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 \ref{corollary:EMD_lower_bound}.

### 3.2 Upper bound part of Theorem \ref{theorem:upper_bound}

Given a known graph \(G_k\) and query access to an unknown graph \(G_u\) (both on \(n\) vertices), we present an algorithm for tolerant testing of graph isomorphism between \(G_k\) and \(G_u\) by using a tolerant EMD tester (for distributions over \(H\)) as a blackbox. Note that this will prove the upper bound part of Theorem \ref{theorem:upper_bound}.

**Algorithm for tolerant GI using tolerant EMD as a black box.** Our testing algorithm is inspired by the algorithm of Fischer and Matsliah \cite{FM08} for non-tolerant GI testing. But our algorithm significantly differs from that of Fischer-Matsliah in some crucial points. As we explain the high level picture of our algorithm, we will point out some of the crucial differences.

We split our algorithm into three phases. In Phase 1, we first choose a \(O\left(\frac{1}{\gamma_2^2 - \gamma_1}\right)\) size collection of random subset of vertices, i.e, \(\text{coresets} C_u\) from the unknown graph \(G_u\) where each \(C_u \in C_u\) is of size \(O(\log n)\). Thereafter we find all embeddings of \(C_u\) inside the known graph \(G_k\). Let the embeddings be \(\eta_1, \eta_2, \ldots, \eta_J\) where \(C^i_k = \eta_i(C_u)\). Now each \(C_u\) (as well as each \(C^i_k\)) 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 \(Y_C^i\)). Now we test if the EMD between \(X_{C_u}\) and \(Y_{C^i_k}\) is close or far for each \(i \in [J]\) (See Claim \ref{claim:EMD_close}). We keep only those \((C_u, \eta_i)\) for Phase 2 such that \(EMD(X_{C_u}, Y_{C^i_k}) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000}) n |C_u|\).

Although Phase 1 of our algorithm is similar to the algorithm of \cite{FM08}, there is a striking difference. Since the authors of \cite{FM08} were testing the non-tolerant version of graph isomorphism, they were testing the identity of the label distributions of \(X_{C_u}\) and \(Y_{C^i_k}\). 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 \textit{good bijections} do not produce much error and testing of tolerant EMD fits exactly for this purpose. It is worth noting that Fischer-Matsliah uses an equivalence tester in their algorithm to identify the placements that do not produce “any” error. But, the proof of correctness of the algorithm would not go through even if we use the tolerant testing of the equivalence of distributions. The use of EMD in this phase is crucial for the proof of correctness of our algorithm to hold.

In Phase 2, we choose \(O\left(\frac{\log^2 n}{(\gamma_2^2 - \gamma_1)}\right)\) 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 all possible labels \(\{l_1, l_2, \ldots, l_t\}\) of the vertices of \(G_k\) under \(C^i_k\)-labelling where \(C^i_k = \eta_i(C_u)\), for those \(\eta_i\) that have passed Phase 1. Ideally, we would like to find a mapping
ψ : W → \{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 ˚φ : W → 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 ∈ 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 ˚φ(W). If this fraction is at most 5γ_1 + 3\frac{γ_2 - γ_1}{5}, we accept and say that G_u and G_k are γ_1-close. If there is no such embedding of any C_u ∈ C_u that achieves this, we report that G_u and G_k are γ_2-far.

The proofs of completeness and soundness follow kind of similar route as Fischer-Matsliah’s proof but the arguments are way more complicated. Many things that were trivial or obvious in the non-tolerant setting become major hurdles in the tolerant setting, and we overcome them with significantly difficult technical arguments, presented in Section 5.

4 Tolerant GI is as hard as tolerant EMD testing

In this section, we prove that it is necessary to perform Ω(QWoR_{EMD}(n)) 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 4.1 (Restatement of the lower bound part of Theorem 1.4). Let G_k be the known and G_u be the unknown graph on n vertices, where n ∈ N is sufficiently large. There exists a constant ε_{ISO} ∈ (0, 1) such that for any given constants γ_1, γ_2 with 0 < γ_1 < γ_2 < ε_{ISO}, any algorithm that decides whether the graphs are γ_1-close or γ_2-far, requires QWoR_{EMD}(n) adjacency queries to the unknown graph G_u where QWoR_{EMD} is as defined in Definition 4.13.

In Section 3.1, we have discussed an overview of our idea to prove the above theorem. To prove Theorem 4.1 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 4.2. Suppose there is a constant ε_0 ∈ (0, \frac{1}{2}) such that for all constants γ_1, γ_2 with 0 < γ_1 < γ_2 < ε_0 and any constant T ∈ 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) ≤ γ_1 T n^2 or d(G_u, G_k) ≥ γ_2 T n^2 by performing Q adjacency queries to G_u.

Then, for any constants β_1 and β_2 with 0 < β_1 < β_2 < \frac{e_0}{2}, the following holds where κ = \frac{β_2 - β_1}{8} and T_κ = \left\lceil\frac{30}{κ(2 - κ)}\right\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_κ n} with |S_k| = |S_u| = n, can distinguish whether EMD(S_k, S_u) ≤ β_1 T κ n^2 or EMD(S_k, S_u) ≥ β_2 T κ n^2 with Q many samples without replacement from S_u.

Remark 4. Observe that Lemma 4.2 talks about tolerant EMD testing between multi-sets with n elements over a Hamming cube of dimension T_κ n. But Theorem 4.1 states the lower bound of QWoR_{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 2.12). So, we will be done with the proof of Theorem 4.1 by proving Lemma 4.2.
4.1 Tolerant GI to Tolerant EMD testing: Proof of Lemma 4.2

To define the necessary reduction for the proof of Lemma 4.2, we need to show the existence of a graph $G_p$ satisfying some unique properties.

**Lemma 4.3.** 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 $n \in \mathbb{N}^{11}$, 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.

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 $(1 - \frac{\kappa}{2})C_{\kappa,s}n$. By using Chernoff bound [2.21] we have

$$\mathbb{P} \left( \deg_{G_p'}(v) \leq ((1-\kappa)C_{\kappa,s} + 1)n \right)$$

$$= \mathbb{P} \left( \deg_{G_p'}(v) \leq (1 - \epsilon) \left(1 - \frac{\kappa}{2}\right)C_{\kappa,s}n \right)$$

$$\leq e^{-\frac{\epsilon^2(2-\kappa)C_{\kappa,s}n}{6}}$$

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

$$\mathbb{P} \left( \mathcal{E}_1 \right) \leq \left| V(G_p') \right| e^{-\frac{\epsilon^2(2-\kappa)C_{\kappa,s}n}{6}} \leq C_{\kappa,s}n \cdot e^{-\frac{\epsilon^2(2-\kappa)C_{\kappa,s}n}{6}}$$

Let $\mathcal{E}_2$ be the event that there exists two (distinct) vertices $u, v$ with $\left| N_{G_p'}(u) \Delta N_{G_p'}(v) \right| < sn$, 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 $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) > 0$.

The rest of the work in this proof is to show $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) > 0$.

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 $C_{\kappa,s} = \lceil \frac{6s}{\kappa(2-\kappa)} \rceil$, $\mathbb{E} \left[ |N_{G_{p}'}(u) \setminus N_{G_{p}'}(v)| \right] \geq 3sn$. Using Chernoff bound \([2.21]\) we have

$$\mathbb{P} \left( |N_{G_{p}'}(u) \setminus N_{G_{p}'}(v)| < sn \right) \leq e^{-\frac{4sn}{9}}$$

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

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

Hence, $\mathbb{P}(\mathcal{E}_{1}^c \cap \mathcal{E}_{2}^c) > 0$.

Let $ALG(\gamma_1, \gamma_2, T)$ be the algorithm that takes $\gamma_1$ and $\gamma_2$ with $0 < \gamma_1 < \gamma_2 < \epsilon_0$ as input and decides whether $d(G_k, G_u) \leq \gamma_1 Tn^2$ or $d(G_k, G_u) \geq \gamma_2 Tn^2$, where $|V(G_k)| = |V(G_u)| = (T + 1)n$.

Now we show that for any two constants $\beta_1$ and $\beta_2$ 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 $(\gamma_1, \gamma_2)$-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.

**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 4.3 with parameters $s = 5$, $\kappa = \frac{\beta_2 - \beta_1}{8}$ and $T_{\kappa} = C_{\kappa,5}$.
- 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 4.3 with parameters $s = 5$, $\kappa = \frac{\beta_2 - \beta_1}{8}$ and $T_{\kappa} = C_{\kappa,5}$. 

18
• Let us first pick a random permutation $\pi$ 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 $j \in [T_n n]$.

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 $A$ 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 $d(G_k, G_u) \leq T_n \gamma_1 n^2$, output that $EMD(S_k, S_u) \leq T_n \beta_1 n^2$. Similarly, if ALG reports that $d(G_k, G_u) \geq T_n \gamma_2 n^2$, then output $EMD(S_k, S_u) \geq T_n \beta_2 n^2$.

**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 4.4** (Special bijections). A bijection $\phi$ 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 $\Phi$, and $d_\Phi(G_k, G_u) := \min_{\phi \in \Phi} d_\phi(G_k, G_u)$.

**Lemma 4.5.** 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 $H_{T_n n} = \{0, 1\}^{T_n n}$. 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 $\psi : S_k \rightarrow S_u$ 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 $\psi'$ and by the definition of $d_{\psi'}(G_k, G_u)$ (See Definition 1), it is clear that $d_{\psi'}(G_k, G_u) = 2 \cdot EMD(S_k, S_u)$. Since $d_{\Phi}(G_k, G_u) = \min_{\psi \in \Phi} d_{\psi}(G_k, G_u)$, 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 $V(G_k) \rightarrow V(G_u)$ that realizes $d_{\Phi}(G_k, G_u)$. By definition of $\Phi$, we can assume that $\psi(b_i) = b_i'$ for each $i \in [T \kappa n]$. Now, let us consider a bijection $\psi'$ 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

$$\sum_{i \in [n]} d_H(k_i, \psi'(k_i)) = \frac{d_{\psi}(G_k, G_u)}{2}.$$  

Thus, $EMD(S_k, S_u) \leq \sum_{i \in [n]} d_H(k_i, \psi'(k_i)) = \frac{d_{\psi}(G_k, G_u)}{2} = \frac{d_{\Phi}(G_k, G_u)}{2}$. Putting everything together, we have $d_{\Phi}(G_k, G_u) = 2 \cdot EMD(S_k, S_u)$. \hfill $\square$

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

**Lemma 4.6.** Let $\Phi$ 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 $d_{\Phi}(G_k, G_u) - 2\kappa T \kappa n^2 \leq d(G_u, G_k) \leq d_{\Phi}(G_k, G_u)$. \footnote{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.}

**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 $\psi : V(G_k) \rightarrow V(G_u)$ that realizes $d(G_k, G_u)$, that is, $d(G_k, G_u) = d_{\psi}(G_k, G_u)$. If $\psi$ is a bijection such that $\psi \in \Phi$, then $d_{\psi}(G_k, G_u) - 2\kappa T \kappa n^2 \leq d(G_k, G_u)$ 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, $d_{\phi}(G_k, G_u) - 2\kappa T \kappa n^2 \leq d(G_u, G_k)$.

We will now present the construction of $\phi \in \Phi$ from $\psi$. Let us first partition the vertices of $B_k$, with respect to $\psi$, into three parts: $B_k = B_{BI} \cup B_{BN} \cup B_A$; 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: $A_k = A_A \cup A_B$; 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 $\psi$) by performing the following three steps in that order. Note that the construction of $\phi$ 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 $b_i \in B_{BI}$, $\phi(b_i) = b_i' \in B_u \setminus \psi(B_{BI})$.

**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, $\phi$ 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 4.3, the graphs $G_k[B_k]$ and $G_u[B_u]$ satisfy the following property: the cardinality of symmetric difference between the sets of neighbors of any two distinct vertices is at least $5n - 2$.

• 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

$$d_\phi(G_u, G_k) \leq d_\psi(G_u, G_k) + 4x(|A_k| + 1) + 2xy + x(x - 1) + 2y |A_k| - y(5n - 2). \quad (2)$$

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

$$d_\phi(G_u, G_k) \leq d_\psi(G_u, G_k) + 4x |A_k| + 4x + 2xy + x(x - 1) + 2y |A_k| - y(5n - 2)$$

$$= d_\psi(G_k, G_u) + 4xn + 4x + 2xy + n(n - 1) + 2ny - y(5n - 2)$$

$$\leq d_\psi(G_k, G_u) + 4n^2 + 4n + 2ny + n^2 + 2ny - y(5n - 2)$$

$$\leq d_\psi(G_k, G_u) + 8n^2$$

$$\leq d_\psi(G_k, G_u) + 2\kappa T \kappa n^2.$$

The last but one inequality follows from the fact that $0 \leq x \leq n$ and the last inequality follows from the fact that $T_\kappa = \lceil \frac{30}{\kappa(2 - \kappa)} \rceil$.

Now we present the proof of Inequality (2).

**Proof of Inequality (2).** Here we prove that

$$d_\phi(G_u, G_k) \leq d_\psi(G_u, G_k) + 4x(|A_k| + 1) + 2xy + x(x - 1) + 2y |A_k| - y(5n - 2). \quad (3)$$

Instead of directly proving the above inequality, we will prove it in four steps for better exposition. In Step 1, we prove the inequality for $x = 1, y = 0$. Then we generalize it for $x \leq n, y = 0$, followed by $x = 0, y \leq T_\kappa n$. Finally, combining Steps 1, 2 and 3, we prove the inequality for any $0 \leq x \leq n$, and $0 \leq y \leq T_\kappa n$.

**Step 1** ($x = 1, y = 0$): So, let us assume that $a_i \in A_k$, $a_j \in A_u$, $b_s \in B_k$ and $b_j \in B_u$ be such that the following holds: $\psi(a_i) = b_s'$ and $\psi(b_s) = a_j'$, $\psi(z) \in A_u$ for each $z \in A_k \{a_i\}$, and $\phi(b_t) = b_t' \in B_u$ for each $b_t \in B_k \{b_s\}$. By the description of Steps (i), (ii) and (iii) of generating $\phi$ from $\psi$, as discussed in Lemma 4.6, we have the following observation.

**Observation 4.7.** 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 $z \in (A_k \cup B_k) \{a_i, b_s\}$, $\phi(z) = \psi(z)$.

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

$$d_\phi(G_k, G_u) \leq d_\psi(G_k, G_u) + 4(|A_k| + 1). \quad (4)$$

By Observation 4.7, $\phi(x) = \psi(x)$ for all vertices $x \in (A_k \cup B_k) \{a_i, b_s\}$. So, any pair of vertices in $(A_k \cup B_k) \{a_i, b_s\}$ has no effect on $d_\phi(G_u, G_k) - d_\psi(G_u, G_k)$. Following Definition 1.1 and Definition 2.2, we can say that

$$d_\phi(G_u, G_k) - d_\psi(G_u, G_k) \leq 2 \left( |\text{DECIDER}_\phi(a_i)| - |\text{DECIDER}_\psi(a_i)| \right. $$

$$\left. + |\text{DECIDER}_\phi(b_s)| - |\text{DECIDER}_\psi(b_s)| \right)$$

Note that we are using Lemma 4.3 with parameters $s = 5$, $\kappa = \frac{2\kappa - \delta_1}{s}$ and $T_\kappa = C_{\kappa, 5}$. 

21
Let us consider $\phi \in \{ \psi \}$. Note that the first term above can be written as $d_\phi(G_u, G_k) = (\text{DECIDER}_\phi(a_i) \cap (A_k \cup \{ b_s \})) \cup (\text{DECIDER}_\phi(a_i) \cap (B_k \setminus \{ b_s \}))$. Breaking other terms in the above expression similarly, we have

\[
d_\phi(G_u, G_k) - d_\psi(G_u, G_k) \\
\leq 2 \left[ 2(|A_k| + 1) + |\text{DECIDER}_\phi(a_i) \cap (B_k \setminus \{ b_s \})| - |\text{DECIDER}_\psi(a_i) \cap (B_k \setminus \{ b_s \})| \\
+ |\text{DECIDER}_\phi(b_s) \cap (B_k \setminus \{ b_s \})| - |\text{DECIDER}_\psi(b_s) \cap (B_k \setminus \{ b_s \})| \right] \\
= 4|A_k| + 4 + 2Z, \text{ where} \\
Z = |\text{DECIDER}_\phi(a_i) \cap (B_k \setminus \{ b_s \})| - |\text{DECIDER}_\psi(a_i) \cap (B_k \setminus \{ b_s \})| \\
+ |\text{DECIDER}_\phi(b_s) \cap (B_k \setminus \{ b_s \})| - |\text{DECIDER}_\psi(b_s) \cap (B_k \setminus \{ b_s \})| \]

By showing $Z \leq 0$, we will be done with the proof of Inequality (4). Observe that we can say $|\text{DECIDER}_\phi(a_i) \cap (B_k \setminus \{ b_s \})| = |\phi(N_{B_k \setminus \{ b_s \}}(a_i)) \Delta N_{B_k \setminus \{ b_s \}}(\phi(a_i))|$. Also, writing the other terms in the expression of $Z$ in the similar fashion, we get

\[
Z \leq |\phi(N_{B_k \setminus \{ b_s \}}(a_i)) \Delta (N_{B_k \setminus \{ b_s \}}(\phi(a_i)))| - |\psi(N_{B_k \setminus \{ b_s \}}(a_i)) \Delta (N_{B_k \setminus \{ b_s \}}(\psi(a_i)))| \\
+ |\phi(N_{B_k \setminus \{ b_s \}}(b_s)) \Delta (N_{B_k \setminus \{ b_s \}}(\phi(b_s)))| - |\psi(N_{B_k \setminus \{ b_s \}}(b_s)) \Delta (N_{B_k \setminus \{ b_s \}}(\psi(b_s)))| \\
\]

Once again, from Observation 4.1.7,

\[
\phi(N_{B_k \setminus \{ b_s \}}(a_i)) = \psi(N_{B_k \setminus \{ b_s \}}(a_i)) \text{ (Say I)} \\
N_{B_k \setminus \{ b_s \}}(\phi(a_i)) = N_{B_k \setminus \{ b_s \}}(\psi(b_s)) \text{ (Say I2)} \\
\phi(N_{B_k \setminus \{ b_s \}}(b_s)) = \phi(N_{B_k \setminus \{ b_s \}}(b_s)) \text{ (Say I3)} \\
N_{B_k \setminus \{ b_s \}}(\psi(a_i)) = N_{B_k \setminus \{ b_s \}}(\phi(b_s)) \text{ (Say I4)} \\
\]

From our above derivation, $|I_2 \Delta I_4| = |\text{DECIDER}_\phi(b_s) \cap (B_k \setminus \{ b_s \})|$. Also, as $y = 0$, we have

\[
|\text{DECIDER}_\phi(b_s) \cap (B_k \setminus \{ b_s \})| = 0. \\
\]

So, to prove $Z \leq 0$, it is enough to show $Z \leq 2|I_3 \Delta T_4|$. Note that

\[
Z \leq |I_1 \Delta I_2| - |I_1 \Delta I_4| + |I_3 \Delta I_4| - |I_3 \Delta I_2|. \\
\]

By using triangle inequality, $Z$ can be upper bounded as follows:

\[
Z \leq |I_2 \Delta I_4| + |I_3 \Delta I_4| - |I_3 \Delta I_2| \leq |I_3 \Delta I_4| + |I_3 \Delta I_4| = 2|I_3 \Delta I_4| = 0. \\
\]

**Step 2 ($x \leq n, y = 0$):** Let us consider $A_B \subseteq A_k$ and $B_A \subseteq B_k$ such that $\psi(a_i) \in B_u$ for each $a_i \in A_B$, $\psi(b) \in B_u$ for each $b \in B_A$, $\psi(a_i) \in A_k$ for each $a_i \in A_k \setminus A_B$, and $\psi(b) \in B_u$ for each $b \in B_k \setminus B_A$. Now let us consider swapping (described below) the mapping of $a_i \in A_B$ and $b \in B_A$ such that $\psi(a_i) = b$. Let $a'_j \in A_u$ be such that $\psi(b) = a'_j$. Let us construct $\phi_x = \psi$ such that the followings hold: $\phi_x(a_i) = a'_j$, $\phi_x(b) = b'_s$, and $\phi_x(\psi(a)) = \psi(\psi(a))$ for each $a \in (A_k \cup B_k) \setminus \{ a_i, b_s \}$. Proceeding in the similar fashion as in the case when $x = 1$ and $y = 0$, we get

\[
d_{\phi_x}(G_u, G_k) - d_\psi(G_u, G_k) \leq 4|A_k| + 4 + 2|I_3 \Delta I_4|, \\
\]

where $|I_3 \Delta I_4| = |\text{DECIDER}_\phi(b_s) \cap (B_k \setminus \{ b_s \})| \leq x - 1$. So,

\[
d_{\phi_x}(G_u, G_k) \leq d_\psi(G_u, G_k) + 4|A_k| + 4 + 2(x - 1). \\
\]

We can proceed in the similar fashion by performing swapping operation of the vertices in $A_B$ and $Y_k$ one by one, and construct $\phi_x = \psi, \phi_{x-1}, \phi_{x-2}, \ldots, \phi_0 = \phi$. Observe that $d_{\phi_{x-1}}(G_u, G_k) \leq d_{\phi_x}(G_u, G_k) + 4|A_k| + 4 + 2(i - 1)$. Also, note that $\phi$ is a special bijection, and moreover

\[
d_\phi(G_u, G_k) \leq 4x|A_k| + 4x + x(x - 1). \\
\]
Step 3 ($x = 0, y \leq T_\kappa n$): Let us consider $B_{BN} \subseteq B_k$ such that $|B_{BN}| = y$. Note that for each $b_s \in B_{BN}$, $\psi(b_s) \neq b'_s$. Consider $b_s \in B_{BN}$ such that $\psi(b_s) = b'_s$, and let $b_j$ be such that $\psi(b_j) = b'_j$. Let us construct $\phi_{y-1} : V(G_u) \rightarrow V(G_k)$ from $\phi_y = \psi$ as follows: $\phi_{y-1}(b_s) = b'_s$, $\phi_{y-1}(b_j) = b'_j$, and $\phi_{y-1}(z) = \psi(z)$ for each $z \in (A_k \cup B_k) \setminus \{b_s, b_j\}$. Thus,

$$d_{\phi_{y-1}}(G_u, G_k) \leq d_{\phi_y}(G_u, G_k) + 2|A_k| - (5n - 2)$$

The term $2|A_k|$ corresponds to the fact that any vertex of $B_{BN}$ has at most $|A_k|$ many neighbors in $A_k$. The second term comes due to the properties of the probabilistic construction of $B_k$ and $B_u$ following Lemma 4.3.

Step 4 ($x \leq n, y \leq T_\kappa n$): Let us assume $\psi(a_i) = b'_s$. Now there are two possibilities:

1. $\psi(b_s) = a'_i$.
2. $\psi(b_s) = b'_i$.

For (1), following the discussion of $x \leq n, y = 0$, we can say that

$$d_{\phi_{x-1}, y} \leq d_{\phi}(G_u, G_k) + 4(|A_k| + 1) + 2(x + y - 1).$$

For (2), we follow the discussion of $x = 0, y \leq T_\kappa n$, and the following holds:

$$d_{\phi_{x-1}}(G_u, G_k) \leq d_{\phi}(G_u, G_k) + 2|A_k| - (5n - 2).$$

Putting everything together, we have

$$d_{\phi}(G_u, G_k) \leq d_{\phi}(G_u, G_k) + 4x(|A_k| + 1) + 2xy + x(x - 1) + 2y |A_y| - y(5n - 2).$$

□

The following lemma completes the proof of Lemma 4.2.

Lemma 4.8. The described algorithm $A$ for EMD, that uses Algorithm ALG on $G_k$ and $G_u$ with parameters $\gamma_1$ and $\gamma_2$ 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 4.5, we have $d_{\Phi}(G_k, G_u) = 2 \cdot EMD(S_k, S_u)$, where $\Phi$ 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 4.4) and Lemma 4.6, 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 4.6, it follows that $d_{\Phi}(G_k, G_u) - 2\kappa T_\kappa n^2 \leq d(G_k, G_u)$. Thus, $d(G_k, G_u) \geq T_\kappa (2\beta_2 - 2\kappa) n^2 = T_\kappa \gamma_2 n^2$.

□
5 Tolerant EMD testing is as hard as tolerant GI

In this section, we prove the following theorem, that discusses about algorithm for tolerant graph isomorphism testing with a blackbox access to tolerant EMD testing over multi-sets.

**Theorem 5.1.** (Restatement of the upper bound part of Theorem 1.4) Let $G_k$ and $G_u$ be the known and unknown graphs, respectively. There exists an algorithm that takes parameters $\gamma_1$ and $\gamma_2$ as input such that $0 \leq \gamma_1 < \gamma_2 \leq 1$, performs $\tilde{O}(\text{QWoREMD}(n))$ many queries to the adjacency matrix of $G_u$ for appropriate $\beta_1$ and $\beta_2$ depending on $\gamma_1$ and $\gamma_2$, 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 $\tilde{O}(\cdot)$ hides a polynomial factor in $\frac{1}{\beta_2 - \beta_1}$ and $\log n$.

**Remark 5.** The theorem stated above works for any $\gamma_1, \gamma_2$ 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 6.** Note that Theorem 5.1 can also be stated in terms of $\text{QWR}_{\text{EMD}}(n)$ as $\text{QWoREMD}(n) \leq \text{QWR}_{\text{EMD}}(n)$ as we can simulate samples with replacement when we have query access to samples without replacement (See Proposition 2.8).

Our algorithm for tolerant GI testing, as stated in Theorem 5.1, 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 $\epsilon_1$ and $\epsilon_2$ 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 5.2.** Let $H = \{0, 1\}^n$ be a $n$-dimensional Hamming cube. Let $\{S^i_k : i \in [t]\} \cup \{S_u\}$ denote the multi-sets with $n$ elements from $H$ where $\{S^i_k : i \in [t]\}$ 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 $\epsilon_1, \epsilon_2$ with $0 \leq \epsilon_1 < \epsilon_2 \leq 1$ and a $\delta \in (0, 1)$ as input and decides whether $\text{EMD}(S_u, S^i_k) \leq \epsilon_1 n^2$ or $\text{EMD}(S_u, S^i_k) \geq \epsilon_2 n^2$, with probability at least $1 - \delta$, for each $i \in [t]$. Moreover, ALG-EMD uses $\text{QWR}_{\text{EMD}}(n) \cdot O(\log \frac{1}{\delta})$ many samples without replacement from $S_u$.

The above theorem follows from the definition of $\text{QWR}_{\text{EMD}}(n)$ (See Definition 1.3) along with union bound and standard argument for amplifying the success probability.

**Remark 7.** The algorithm of Theorem 5.1 to be discussed in Section 5.1, formulates a tolerant $EMD$ instance of multi-sets having $n$ elements in $H = \{0, 1\}^d$, where $d = O(\log n/(\gamma_2 - \gamma_1))$. 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 2.12 in Section 2.3). Moreover, the algorithm in Section 5.1 calls ALG-EMD with parameters $\epsilon_1 = (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000})$, $\epsilon_2 = \gamma_2/5$, $t = 2^{O(\log^2 n/(\gamma_2 - \gamma_1))}$ and $\delta$ is a suitable constant depending upon $\gamma_1$ and $\gamma_2$, where $\gamma_1$ and $\gamma_2$ are parameters as stated in Theorem 5.1. So, each call to ALG-EMD, in our context, makes $\tilde{O}(\text{QWoREMD}(n))$ many queries.

5.1 Algorithm for tolerant GI testing

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

**Definition 5.3.** (Label of a vertex) Given a graph $G$ and $C \subset V(G) = \{c_1, \ldots, c_{|C|}\}$, the $C$-labelling of $V(G)$ is a function $\mathcal{L}_C : V(G) \rightarrow \{0, 1\}^{|C|}$ 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 $V(G)$. 

24
**Definition 5.4.** (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 $\eta$ from $A$ to $B$ is referred to as an embedding of $A$ into $B$.

Now we present our query algorithm $\text{TolerantGI}(G_u, G_k, \gamma_1, \gamma_2)$ that comprises three phases. A he technical overview of the algorithm is already presented in Section 3.2.

**Formal Description of TolerantGI($G_u, G_k, \gamma_1, \gamma_2$):**

The three phases of our algorithm are as follows:

**5.1.1 Phase 1**

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

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

**Step 2** We determine all possible embeddings, that is, $\eta_1, \ldots, \eta_J$ of $C_u$ into $V(G_k)$, where $J = \frac{n}{O(log n)} \leq 2^{O(log^2 n)}$. For each $i \in [J]$, let $C^i_k$ be the set of images of $C_u$ under the $i$-th embedding of $C_u$ into $V(G_k)$, that is, $C^i_k = \eta_i(C_u)$. For all $i \in [J]$, we construct the multi-set $Y_{C^i_k}$ that contains $C^i_k$-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 vertices of $V(G_u)$. However, $X_{C_u}$ is unknown to the algorithm. We call $\text{ALG-EMD}$ (as stated in Theorem 5.2) by setting parameters as described in Remark 7 to decide whether $\text{EMD}(X_{C_u}, Y_{C^i_k}) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000})n|C_u|$ or $\text{EMD}(X_{C_u}, Y_{C^i_k}) \geq \gamma_2 n |C_u|/5$, for each $i \in [J]$. Let us pair up $C_u$'s and their accepted embeddings into $G_k$ and call the set $\Gamma$, that is,

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

Note that, at the end of the **Phase 1**, we have $|\Gamma| \leq |C_u| \cdot 2^{O(log^2 n)} = O\left(2^{(log^2 n)}\right)$. By the description of **Step 3** above, **Phase 1** of our algorithm calls $\text{ALG-EMD}$ $O(|C_u|)$ times, once for each $C_u \in C_u$. So, setting $\delta = \frac{1}{2000}$ in Theorem 5.2, we obtain the following observation about $\Gamma$ that will be used to prove the soundness of our algorithm.

**Observation 5.5.** Consider $\Gamma$, the set of accepted embeddings that have passed **Phase 1** paired with corresponding $C_u$, as defined above. Then

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

**5.1.2 Phase 2**

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

**Step 1** We sample a subset $W$ of $O(\log^2 n/(\gamma_2 - \gamma_1)^3)$ 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:
We randomly pair up the vertices of $G_k$. Let $l_1, \ldots, l_t$ be the labels of the vertices where $t = 2^{|C_k|}$ and $V_j \subseteq V(G_k)$ be the set of vertices with label $l_j$.

(ii) We define a matrix $M$ of size $|W| \times 2^{|C_k|}$ where each row represents the label of a vertex $w \in W$ and each column represents one of the possible $C_k$-labelling of $V(G_k)$ \footnote{Let $C_k = \{x_1, \ldots, x_{\Omega(\log n/(\gamma_2 - \gamma_1))}\}$. Note that for each $w \in W$, $\mathcal{L}_{C_k}(w_i) \in \{0, 1\}^{\Omega(\log n/(\gamma_2 - \gamma_1))}$ such that the $j$-th coordinate is 1 if and only if $w_i$ is a neighbour of $x_j$, where $i \in [\Omega(\log n/(\gamma_2 - \gamma_1))]$ and $j \in [\Omega(\log n/(\gamma_2 - \gamma_1))]$. Similarly, $l_j \in \{0, 1\}^{\Omega(\log n/(\gamma_2 - \gamma_1))}$ 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 $j \in [2^{|C_k|}]$.} The $(i, j)$-th entry of $M$ is defined as: $M_{ij} = d_H(\mathcal{L}_{C_k}(w_i), l_j)$.

(iii) We choose a function $\psi : W \to \{l_1, \ldots, l_t\}$ randomly satisfying

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

Let $\Gamma_W$ be the set of tuples such that

$$\Gamma_W = \{(C_u, \eta_i, \psi) : (C_u, \eta_i) \in \Gamma \text{ and } \psi \text{ satisfies Equation } (5)\}.$$ 

Like Observation 5.5, the following observation about the set $\Gamma_W$ will be used to prove the soundness of our algorithm.

**Observation 5.6.** $|\Gamma_W| \leq |\Gamma| \leq 2^{O(\log^2 n)}$. Moreover, any $(C_u, \eta_i, \psi)$ that has passed this phase satisfies Equation (5).

### 5.1.3 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 $p = O(\log^2 n/(\gamma_2 - \gamma_1)^3)$. 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 $\hat{\psi} : W \to V(G_k)$ randomly, satisfying $\hat{\psi}(w) \in V_j$ 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 $C_k$-labelling in $G_k$.

**Step 4** We find the fraction $\zeta(C_u, \eta_i, \psi, \hat{\psi}) = \frac{|\{(a_i, b_i) : 1_{(a_i, b_i)} = 1\}|}{p}$, where $1_{(a_i, b_i)} = 1$ 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 $\zeta(C_u, \eta_i, \psi, \hat{\psi}) \leq 5\gamma_1 + \frac{3}{8}(\gamma_2 - \gamma_1)$, then **HALT and REPORT** that $G_u$ and $G_k$ are $\gamma_1$-close.

While executing Step 3 and Step 4 for each tuple in $\Gamma_W$, if we did not **HALT**, then we **HALT** now and **REPORT** that $G_u$ and $G_k$ are $\gamma_2$-far.

**Observation 5.7.** (i) The number of times our algorithm executes Step 2, Step 3 and Step 4 is at most $|\Gamma_W| \leq 2^{O(\log^3 n)}$.

(ii) If there exists a $(C_u, \eta_i, \psi)$ such that $\zeta(C_u, \eta_i, \psi, \hat{\psi}) \leq 5\gamma_1 + \frac{3}{8}(\gamma_2 - \gamma_1)$, then our algorithm reports that $G_u$ and $G_k$ are $\gamma_1$-close. Otherwise, $G_u$ and $G_k$ are reported to be $\gamma_2$-far.
5.2 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 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 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 $O(n)$.

5.2.1 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 $C_u \in C_u$ 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 $\text{EMD} \left( X_{C_u}, Y_{C_k} \right) \leq (1 + \frac{2\gamma_1}{2000})n |C_u|$ holds with probability at least 20/21, where $C_k^i = \eta_i(C_u)$.

Claim 5.8. 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 $C_u \in C_u$ 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} \right) \leq (1 + \frac{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 $V(G_k)$.

Proof. Consider a particular $C_u \in C_u$ 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 $\eta_i$ is considered in Step 2 of Phase 1 of the algorithm. Now we will show that $\text{EMD} \left( X_{C_u}, Y_{C_k} \right) \leq (1 + \frac{2\gamma_1}{2000})n |C_u|$ holds with probability at least a constant, to be specified later, that depends upon $\gamma_1$ and $\gamma_2$, where $C_k^i = \eta_i(C_u)$.

We know that $d_\phi(G_u, G_k) \leq \gamma_1 n^2$ and by Definition 2.2, we have

$$\sum_{x \in V(G_u)} |\text{DECIDER}_\phi(x)| \leq \gamma_1 n^2.$$  

Thus,

$$E \left[ \sum_{x \in V(G_u)} |\text{DECIDER}_\phi(x) \cap C_u| \right] \leq \gamma_1 n |C_u|.$$  

(6)

From Definition 2.2, we can say that

$$\text{EMD} \left( X_{C_u}, Y_{C_k} \right) = \min_{f : V(G_u) \to V(G_k)} \sum_{x \in V(G_u)} |\text{DECIDER}_f(x) \cap C_u|$$

$$\leq \sum_{x \in V(G_u)} |\text{DECIDER}_\phi(x) \cap C_u|$$

$^{15}$C_k^i and Y_{C_k^i} are defined in Step 2 of Phase 1.
Therefore,

\[
\mathbb{E} \left[ \text{EMD} \left( X_{C_u}, Y_{C_k^i} \right) \right] \leq \mathbb{E} \left[ \sum_{x \in V(G_u)} |\text{DECIDER}_\phi(x) \cap C_u| \right] \\
\leq \gamma_1 n \left| C_u \right| \quad \text{(From Equation (6))}
\]

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 \( |C_u| = O(\frac{1}{\gamma_2 - \gamma_1}) \) and we have been arguing for a particular \( C_u \in C_u \). So, taking \( |C_u| \) suitably, we get a \( C_u \) and an embedding \( \eta_i : C_u \rightarrow V(G_k) \) 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 \( C_u \in C_u \) and its embeddings satisfying above mentioned desired properties. Now we discuss how our algorithm determines all \( C_u \in C_u \) that satisfy the properties. Note that Step 3 of Phase 1 of our algorithm calls ALG-EMD. Following the correctness of ALG-EMD (Theorem 5.2), we determine all embeddings \( \eta_i : C_u \rightarrow V(G_k) \) 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 5.9. Let \( C_u \in C_u \) 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 \( \Gamma = \{(C_u, \eta_i) \mid \text{EMD}(X_{C_u}, Y_{C_k^i}) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000}) n \left| C_u \right| \} \) 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 \( \gamma_1 \)-close to being isomorphic, from Claim 5.8 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 \( \psi : W \rightarrow \{l_1, \ldots, l_t\} \) as considered in Step 2 (iii) in Phase 2 of our algorithm such that Equation (5) holds with probability at least 20/21.

Claim 5.10. Let us assume that \( \phi : V(G_u) \rightarrow V(G_k) \) be a bijection such that \( d_\phi(G_u, G_k) \leq \gamma_1 n^2 \) and \( (C_u, \eta_i) \in \Gamma \) where \( C_u \in C_u \) and \( \eta_i : C_u \rightarrow V(G_k) \) be an embedding such that

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

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

\[
(i) \sum_{w \in W} d_H(C_{C_u}(w), \psi(w)) \leq \frac{2\gamma_2}{\gamma} |C_u||W|, \quad \text{and} \\
(ii) \forall j \in [t], \text{ we have } |\{w : \psi(w) = l_j\}| \leq |V_j|.
\]

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

28
Since \(|\text{DECIDER}_f(x) \cap C_u| = d_H(\mathcal{L}_{C_u}(x), \mathcal{L}_{C_k}(f(x)))\), we have
\[
\sum_{x \in V(G_u)} d_H(\mathcal{L}_{C_u}(x), \mathcal{L}_{C_k}(f(x))) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000}) n |C_u|
\]
Since we are taking the vertices in \(W\) uniformly at random from \(G_u\), we can say that
\[
\mathbb{E} \left[ \sum_{w \in W} d_H(\mathcal{L}_{C_u}(w), \mathcal{L}_{C_k}(f(w))) \right] \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000}) |C_u| |W|
\]
Using Hoeffding’s inequality, we have
\[
\mathbb{P} \left( \sum_{w \in W} d_H(\mathcal{L}_{C_u}(w), \mathcal{L}_{C_k}(f(w))) \leq \frac{2\gamma_2}{5} |C_u| |W| \right) \geq 1 - e^{-O(|W|)}
\]
Now, we define \(\psi : W \to \{\ell_1, \ldots, \ell_t\}\) such that \(\psi(w) = \mathcal{L}_{C_k}(f(w))\). In other words, the \(C_k\)-labelling of \(f(w)\) is same as the labelling of \(\psi(w)\) for each \(w \in W\). Thus, the \(\psi\) 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
\[
|\{w \in W : \mathcal{L}_{C_k}(f(w)) = l_j\}| \leq |\{v \in V(G_k) : \mathcal{L}_{C_k}(v) = l_j\}| \leq |V_j|.
\]
So, by the definition of \(\psi\), \(|\{w \in W : \psi(w) = l_j\}| \leq |V_j|\). Hence \(\psi\) 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 \(\gamma_1\)-close, that is, there exists a bijection \(\phi\) from \(V(G_u)\) to \(V(G_k)\) such that \(d_\phi(G_u, G_k) \leq \gamma_1 n^2\), then there exists \(C_u \subset C_u, \eta_i : C_u \to V(G_k),\) and \(\psi\) satisfying the conditions given in Claims 5.8 and 5.10. However, we do not know \(\phi\). If we construct, though inefficiently, a bijection \(\phi'\) that is same as \(\phi\) with respect to the same \(C_u \subset C_u, \eta_i : C_u \to V(G_k)\) and \(\psi\) (conditions given in Claims 5.8 and 5.10), 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 5.11.** 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 \(C_u \subset C_u\) and \(\eta_i : C_u \to V(G_k)\) be an embedding such that

\begin{itemize}
  \item \(\forall v \in C_u\) we have \(\eta_i(v) = \phi(v)\), and
  \item \(\text{EMD}\left(X_{C_u}, Y_{C_k}\right) \leq (\gamma_1 + \frac{\gamma_2 - \gamma_1}{2000}) n |C_u|\) where \(C_k = \eta_i(C_u)\).
\end{itemize}

Let \(\{\ell_1, \ldots, \ell_t\}\) be the all possible \(C_k\)-labellings of the vertices of \(G_k\) where \(t = \left\lceil \frac{2|C_k|}{\gamma_2} \right\rceil\), 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

\begin{itemize}
  \item \(\sum_{w \in W} d_H(\mathcal{L}_{C_u}(w), \psi(w)) \leq \frac{2\gamma_2}{5} |C_u| |W|,\) and
  \item \(\forall j \in [t],\) we have \(|\{w : \psi(w) = l_j\}| \leq |V_j|\).
\end{itemize}
Then, with probability at least $18/21$, there exists a bijection $\phi' : V(G_u) \to V(G_k)$, with $\phi'(x) = \eta_i(x)$ for each $x \in C_u$ and $\phi'(w) = \hat{\phi}(w)$ for each $w \in W$ such that

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

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

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

(7)

By using Definition 2.2, we write the above equation as

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

So,

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

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

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

(8)

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

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

Before proceeding further, consider the following observation, which follows from standard Chernoff bound type argument.

**Observation 5.12.** If $|\text{Symm}_{\phi'}(x)| \geq \frac{(\gamma_2 - \gamma_1)n}{1000}$, then

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

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

$$\sum_{x \in V(G_u): |\text{Symm}_{\phi'}(x) \cap C_u| \geq \frac{(\gamma_2 - \gamma_1)n}{1000}} |\text{Symm}_{\phi'}(x)| \geq \left(1 - \frac{1}{50}\right)\frac{|C_u|}{n} \sum_{x \in V(G_u): |\text{Symm}_{\phi'}(x)| \geq \frac{(\gamma_2 - \gamma_1)n}{1000}} |\text{Symm}_{\phi'}(x)|$$

$$= \frac{49}{50} \left(4\gamma_1 + \frac{499(\gamma_2 - \gamma_1)}{1000}\right) n |C_u|. \text{ (\because By Equation (8))}$$

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

$$\sum_{x \in V(G_u)} |\text{Symm}_{\phi'}(x) \cap C_u| \geq \frac{49}{50} \left(4\gamma_1 + \frac{499(\gamma_2 - \gamma_1)}{1000}\right) n |C_u|. \quad (9)$$
Assuming Equation (9) holds and using the fact that $W \subset V(G_u)$ is taken uniformly at random, we can say that

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

Using Hoeffding’s inequality (See Lemma 2.23), we get

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

As the above equation holds in the conditional space that Equation (9) holds, we have

$$\mathbb{P} \left( \sum_{w \in W} |\text{Symm}_{\phi'}(w) \cap C_u| > (3\gamma_1 + \frac{11(\gamma_2 - \gamma_1)}{24}) |C_u| |W| \right) \geq 1 - ne^{-O(|C_u|)} - e^{-O(|W|)}.$$  \hfill (10)

Note that Equation (7) implies Equation (10). However, till now, we have not used any information given in the statement of Claim 5.11 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 $\psi$ is bounded, we will deduce that

$$\mathbb{P} \left( \sum_{w \in W} |\text{Symm}_{\phi'}(w) \cap C_u| \leq (2\gamma_1 + \frac{9(\gamma_2 - \gamma_1)}{20}) |C_u| |W| \right) \geq 1 - ne^{-O(|C_u|)} - e^{-O(|W|)}.$$  \hfill (11)

As Equation (7) implies Equation (10), and Equations (10) and (11) together implies that Equation (7) does not hold with probability at least $1 - 4ne^{-O(|C_u|)} - e^{-O(|W|)}$. Hence, we are done with the proof of Claim 5.11 except that we need to show Equation (11).

By the definition of the bijection $\phi$, we have $\sum_{x \in V(G_u)} |\text{DECIDER}_\phi(x)| \leq \gamma_1 n^2$. This implies

$$\sum_{x \in V(G_u)} |\text{DECIDER}_\phi(x)| \leq \gamma_1 n^2$$ \hfill (12)

To proceed further, we need the following observation, which is a direct application of Chernoff-Hoeffding bound.

**Observation 5.13.** (i) If $|\text{DECIDER}_\phi(x)| \geq (\frac{\gamma_2 - \gamma_1}{} n)$, then

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

(ii) If $|\text{DECIDER}_\phi(x)| < (\frac{\gamma_2 - \gamma_1}{} n)$, then $\mathbb{P} \left( |\text{DECIDER}_\phi(x) \cap C_u| \geq \frac{\gamma_2 - \gamma_1}{150} |C_u| \right) \leq e^{-O(|C_u|)}.$

Note that the above observation implies that the following holds with probability at least $1 - ne^{-O(|C_u|)}.$
Let us assume Equation (13) holds. Since we are taking the vertices of \( V(G_u) \) uniformly at random from \( V(G_u) \), we have

\[
\sum_{x \in V(G_u)} |\text{DECIDER}_\phi(x) \cap C_u| = \sum_{x \in V(G_u): |\text{DECIDER}_\phi(x)| \geq \frac{(\gamma_2 - \gamma_1) n}{1000}} |\text{DECIDER}_\phi(x) \cap C_u| + \sum_{x \in V(G_u): |\text{DECIDER}_\phi(x)| < \frac{(\gamma_2 - \gamma_1) n}{1000}} |\text{DECIDER}_\phi(x) \cap C_u| 
\leq \left(1 + \frac{1}{50}\right) \sum_{x \in V(G_u): |\text{DECIDER}_\phi(x)| \geq \frac{(\gamma_2 - \gamma_1) n}{1000}} |\text{DECIDER}_\phi(x)| \frac{|C_u|}{n} + \frac{\gamma_2 - \gamma_1 n |C_u|}{750} \leq \frac{51}{50} \gamma_1 n |C_u| + \frac{\gamma_2 - \gamma_1 n |C_u|}{750}.
\]

Note that the last inequality follows from Equation (12). Summarizing the above calculation, we get that the following event occurs with probability at least \( 1 - ne^{-\Theta(C_u)} \).

\[
\sum_{x \in V(G_u)} |\text{DECIDER}_\phi(x) \cap C_u| \leq \frac{51}{50} \gamma_1 n |C_u| + \frac{\gamma_2 - \gamma_1 n |C_u|}{750}. \tag{13}
\]

Let us assume Equation (13) holds. Since we are taking the vertices of \( W \) uniformly at random from \( V(G_u) \), we have

\[
\mathbb{E} \left[ \sum_{w \in W} |\text{DECIDER}_\phi(w) \cap C_u| \right] = \mathbb{E} \left[ \sum_{w \in W} d_H(\mathcal{L}_{C_{u}}, \mathcal{L}_{C_{u}^k}(\phi(w))) \right] \leq \frac{51}{50} \gamma_1 |C_u| |W| + \frac{\gamma_2 - \gamma_1 |C_u| |W|}{750}.
\]

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

\[
\sum_{w \in W} |\text{DECIDER}_{\phi'}(w) \cap C_u| = \sum_{w \in W} d_H(\mathcal{L}_{C_{u}}, \mathcal{L}_{C_{u}^k}(\phi'(w))) \leq \frac{2 \gamma_2}{5} |C_u| |W|
\]

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

\[
\mathbb{E} \left[ \sum_{w \in W} |\text{Symm}_{\phi\phi'}(x) \cap C_u| \right] \leq \mathbb{E} \left[ \sum_{w \in W} |\text{DECIDER}_{\phi'}(w) \cap C_u| \right] + \mathbb{E} \left[ |\text{DECIDER}_{\phi}(w) \cap C_u| \right] \leq \left( \frac{764}{750} \gamma_1 + \frac{301(\gamma_2 - \gamma_1)}{750} \right) |C_u| |W|
\]

Using Hoeffding’s inequality (see Lemma 2.23), we can say that

\[
P \left( \sum_{w \in W} |\text{Symm}_{\phi\phi'}(w) \cap C_u| > \left(2 \gamma_1 + \frac{9(\gamma_2 - \gamma_1)}{20} \right) |C_u| |W| \right) \leq e^{-\Theta(C_u^2 |W| |C_u|)} = e^{-\Theta(|W|)}.
\]

32
If we had constructed a bijection $\phi'$ 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 $\phi'$ efficiently. However, without constructing the bijection $\phi'$, 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 $\hat{\phi} : W \rightarrow V(G_k)$ is the map constructed in Step 3 of Phase 3 and $\phi' : V(G_u) \rightarrow V(G_k)$ is the bijection as stated in Claim 5.11. 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 $\gamma_1$-close, then $\zeta(C_u, \eta, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)$, as considered in Step 4 of Phase 3 holds with probability at least 20/21.

**Claim 5.14.** Let us assume that $\phi : V(G_u) \rightarrow V(G_k)$ be a bijection such that $d_\phi(G_u, G_k) \leq \gamma_1 n^2$, and $(C_u, \eta_v) \in \Gamma$ where $C_u \in C_u$, and $\eta : C_u \rightarrow V(G_k)$ be an embedding of $C_u$ such that

- $\forall v \in C_u$ we have $\eta_v(v) = \phi(v)$, and
- $\text{EMD} \left( X_{C_u}, Y_{C_u} \right) \leq (\gamma_1 + \frac{\sqrt{2} - 2}{2000})n |C_u|$ where $C_u^{\phi} = \eta(C_u)$.

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

- $\sum_{w \in W} d_H(L_{C_u}(w), \psi(w)) \leq \frac{2\gamma_2}{n} |C_u||W|$, and
- $\forall j \in \ell_t$, we have $|\{w : \psi(w) = \ell_j\}| \leq |V_j|$.

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

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

holds with probability at least 20/21, where $\zeta(C_u, \eta, \psi, \hat{\phi})$ 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 $\hat{\phi} : W \rightarrow V(G_k)$ in Step 4 of Phase 3 to compute $\zeta(C_u, \eta, \psi, \hat{\phi})$. Considering the conditions given in the statement of this claim and Claim 5.11, 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 $\phi' : V(G_u) \rightarrow V(G_k)$, where $\phi'$ is a bijection with $d_{\phi'}(G_u, G_k) \leq (5\gamma_1 + \frac{3\gamma_2 - \gamma_1}{2})n^2$.

So, $\mathbb{E}\left[\zeta(C_u, \eta_i, \psi, \phi)\right] \leq (5\gamma_1 + \frac{3\gamma_2 - \gamma_1}{2})$. Now, applying Hoeffding’s inequality (Lemma 2.23) and taking $|W| = C' \frac{\log^2 n}{(\gamma_2 - \gamma_1)^2}$ for suitably large constant $C'$, we have

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

\[\square\]

Now we are ready to prove the completeness property using Claims 5.8, 5.10, 5.11, 5.14 and Theorem 5.2.

**Lemma 5.15** (Completeness Lemma). If $G_u$ and $G_k$ are $\gamma_1$-close to isomorphic, then our algorithm reports the same with probability at least 2/3.

**Proof.** Observe that from Claim 5.8, we know that, with probability at least 20/21, there exists a $C_u \in C_u$ and an embedding $\eta_i : C_u \rightarrow V(G_k)$ such that $\text{EMD}(X_{C_u}, Y_{C_k}) \leq \left(\gamma_1 + \frac{3\gamma_2 - \gamma_1}{2000}\right)n |C_u|$ where $C_k = \eta_i(C_u)$. Similarly, from Theorem 5.2 we can say that, with probability at least 20/21, the algorithm ALG-EMD returns all embeddings $\eta_i$ such that $\text{EMD}(X_{C_u}, Y_{C_k}) \leq \left(\gamma_1 + \frac{3\gamma_2 - \gamma_1}{2000}\right)n |C_u|$.

Now from Claim 5.10, we know that, with probability at least 20/21, conditions of Equation 5 hold. Again, from Claim 5.11, we can say that constructing partial bijection at Step 3 of Phase 3 does not change isomorphism distance by more than $(4\gamma_1 + \frac{3\gamma_2 - \gamma_1}{2})n^2$ with probability at least 18/21. Finally, from Claim 5.14, we can say that the algorithm will correctly detect the distance at Step 4 of Phase 3 by testing $\zeta(C_u, \eta_i, \psi, \phi') \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)$ with probability at least 20/21. Thus, using union bound, we can say that when $G_k$ and $G_u$ are $\gamma_1$-close to being isomorphic, TolerantGI($G_u$, $G_k$, $\gamma_1$, $\gamma_2$) reports the same with probability at least 2/3. \[\square\]

### 5.2.2 Proof of Soundness Property

Similarly for the soundness property of our algorithm, let us consider the case when $G_u$ and $G_k$ are $\gamma_2$-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 $\Gamma_W$ with which we started Phase 3 of our algorithm.

$$\Gamma_W = \{(C_u, \eta_i, \psi) : (C_u, \eta_i) \in \Gamma \text{ such that Equation } 5 \text{ holds}\}.$$  

By Observation 5.5, we have

$$\text{Pr}\left(\forall (C_u, \eta_i, \psi) \in \Gamma_W, \text{EMD}(X_{C_u}, Y_{C_k}) \leq \frac{\gamma_2}{5}|C_u|n\right) \geq \frac{8}{9}. \quad (14)$$

From now on, we work on the conditional space where $\text{EMD}(X_{C_u}, Y_{C_k}) \leq \frac{3\gamma_2}{5}|C_u|n \forall (C_u, \eta_i, \psi)$ holds. By Observation 5.7(i), we know that $|\Gamma_W| \leq 2^O(\log^2 n/(\gamma_2 - \gamma_1))$. So, the following claim about any $(C_u, \eta_i, \psi) \in \Gamma_W$ along with union bound over all the elements in $\Gamma_W$, we will be done with the proof of soundness property.
Claim 5.16. 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 \((C_u, \eta_i, \psi)\). Also, let \(\text{EMD}(X_G, Y'_G) \leq \frac{2\gamma_2}{5} |C_u| n\), where \(C^k_u = \eta_i(C_u)\). Then the following holds with probability at most \(\frac{2}{9|\Gamma_W|}\):

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

Proof. Let \(\Phi(C_u, C^k_u)\) 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
- \(\sum_{v \in V(G_u)} \text{DECIDER}_\phi(v) \cap C_u \leq \frac{2\gamma_2}{5} n |C_u|\).

Consider the following observation, about the bijections in \(\Phi\), that we will prove later.

Observation 5.17. Let \(\phi\) be a bijection in \(\Phi\). Then \(\sum_{w \in W} |\text{DECIDER}_\phi(w) \cap C_u| \leq \frac{2\gamma_2}{5} |C_u| |W|\) holds with probability at least \(1 - \frac{1}{9|\Gamma_W|}\).

Our algorithm constructs \(\psi : W \to \{\ell_1, \ldots, \ell_t\}\) in Step 2 of Phase 2 satisfying

- \(\sum_{w \in W} d_H(L_{C_u}(w), \psi(w)) \leq \frac{2\gamma_2}{5} |C_u| |W|\), and
- \(\forall j \in [t], we have \{|w : \psi(w) = l_j\}| \leq |V_j|\).

Note that \(\sum_{w \in W} d_H(L_{C_u}(w), \psi(w)) = \sum_{w \in W} |\text{DECIDER}_\phi(w) \cap C_u|\), where \(\phi\) is some bijection in \(\Phi\). After getting \(\psi\), 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 \(\sum_{w \in W} |\text{DECIDER}_\phi(w) \cap C_u| \leq \frac{2\gamma_2}{5} |C_u| |W|\). Now, from Observation 5.17, we have the following observation.

Observation 5.18. \(\hat{\phi}\) 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 \(\phi\) such that \(\phi|_W = \hat{\phi}\). Let \(W = \{\hat{\phi}_X = \phi|_X : X \subset V(G_u)\ and \ |X| = |W|\}\), and \(W' \subseteq W\) is defined as:

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

Observe that \(\hat{\phi} = \hat{\phi}_W \in W\). By Observation 5.17, we know that if we take a set \(X \subset V(G_u)\) (i.e., a \(\hat{\phi}_X\) uniformly at random from \(W\)), then the probability that \(\hat{\phi}_X \in W'\), is at least \(1 - \frac{1}{9|\Gamma_W|}\). So, \(|W'| \geq \left(1 - \frac{1}{9|\Gamma_W|}\right) |W|\).

Observe that the partial bijection \(\hat{\phi}\), constructed by our algorithm, is same as that of \(\hat{\phi}_W\), and \(\hat{\phi}\) is in \(W'\). Now, using the fact that \(|W'| \geq \left(1 - \frac{1}{9|\Gamma_W|}\right) |W|\), 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 \( \frac{|W|}{3} \) 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, \( \mathbb{E}\left[ \zeta(C, \eta, \psi, \hat{\phi}) \right] \geq \gamma_2 |W| \). Now we can deduce the following. \(^{16}\)

\[
\mathbb{P}\left( \zeta(C, \eta, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1) \right) = \mathbb{P}\left( \zeta(C, \eta, \psi, \hat{\phi}) |W| \leq (5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1)) |W| \right) \\
\leq e^{-O(|W|)} \\
\leq \frac{1}{9 |W|}
\]

Note that we were deriving the above bound on \( \mathbb{P}\left( \zeta(C, \eta, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1) \right) \) assuming that \( \hat{\phi} \) is a random restriction of a random \( \phi \in \Phi \). Hence, combining Observation 5.18 with the above bound on \( \mathbb{P}\left( \zeta(C, \eta, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1) \right) \) (when \( \hat{\phi} \) is a random restriction of a random \( \phi \in \Phi \)), we get

\[
\mathbb{P}\left( \zeta(C, \eta, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1) \right) \leq \frac{2}{9 |W|}.
\]

\[
\text{Proof of Observation 5.17}. \text{ Since } W \text{ is taken uniformly at random,}
\]

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

Using Hoeffding’s inequality, we get

\[
\mathbb{P}\left( \sum_{w \in W} |\text{DECIDER}_\phi(w) \cap C_u| \geq \frac{2\gamma_2}{5} |C_u| |W| \right) \leq e^{-O(|W|)} \leq \frac{1}{9 |W|}.
\]

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

**Lemma 5.19 (Soundness Lemma).** If \( G_u \) and \( G_k \) are \( \gamma_2 \)-far from isomorphic, then the algorithm reports the same with probability at least 2/3.

\[
\text{Proof.} \text{ From Observation 5.7 (i), we know that } |\Gamma_W| \text{ is at most } 2^C_1 \frac{\log^2 n}{\gamma_2 - \gamma_1}. \text{ In Claim 5.16, we are proving that } \zeta(C, \eta, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1) \text{ holds with probability at most } \frac{2}{9 |W|} \text{ for any particular } (C, \eta, \psi) \in \Gamma_W \text{ with } EMD(X_{C_u}, Y_{C_k^\phi}) \leq \frac{\gamma_2}{5} |C_u| n. \text{ So, by the union bound, the probability that there exists a } (C, \eta, \psi) \in \Gamma_W \text{ with } EMD(X_{C_u}, Y_{C_k^\phi}) \leq \frac{\gamma_2}{5} |C_u| n \text{ such that } \zeta(C, \eta, \psi, \hat{\phi}) \leq 5\gamma_1 + \frac{3}{5}(\gamma_2 - \gamma_1), \text{ is at most } \frac{2}{5}. \text{ Now From Equation } 14
\]

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

Putting everything together, the probability that the algorithm reports that \( G_u \) and \( G_k \) are \( \gamma_2 \)-far, is at least 2/3.

\(^{16}\)Here we are assuming \( \gamma_2 \geq 11\gamma_1 \).
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.

5.3 Proof of Theorem 5.1

Proof of Theorem 5.1. From the Completeness Lemma (Lemma 5.15) and Soundness Lemma (Lemma 5.19), 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 \( \text{Cost}_{C_u} \) denote the query complexity corresponding to a particular \( C_u \in \mathcal{C}_u \). So, the total query complexity of the algorithm TolerantGI is \( \sum_{C_u \in \mathcal{C}_u} \text{Cost}_{C_u} \). Observe that

\[
\text{Cost}_{C_u} = \text{Query Complexity of algorithm ALG-EMD} + \text{Cost}_{C_u,W}
\]

where \( \text{Cost}_{C_u,W} \) 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 5.2. In Step 3 of Phase 1 of our algorithm, for each \( C_u \in \mathcal{C}_u \), we call ALG-EMD with parameters \( d = \Theta(\log n) \), \( t = 2^O(\log^2 n) \), \( \epsilon_1 = (\gamma_1 + \frac{2\gamma_2}{2000}) \), \( \epsilon_2 = \frac{\gamma_2}{T} \) and \( \delta = \Theta(1) \). So, the query complexity of each call, to ALG-EMD from our algorithm, is \( \tilde{O}(\min\{n, 2^d\}) = \tilde{O}(n) \).

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

6 Communication Complexity Landscape of GI

Let us first recall the model as stated in Section 2.4. Two players Alice and Bob have two graphs \( G_a \) and \( G_b \) (on \( n \) vertices) respectively. They would like to communicate among themselves to decide about the following problems:

1. **Non-tolerant Graph Isomorphism:** If \( G_a \) and \( G_b \) are isomorphic or \( \epsilon \)-far from isomorphic where \( \epsilon \in (0, 1] \) is a proximity parameter.

2. **Tolerant Graph Isomorphism:** If \( G_a \) and \( G_b \) are \( \epsilon_1 \)-close to being isomorphic or \( \epsilon_2 \)-far from being isomorphic where \( \epsilon_1, \epsilon_2 \) are two proximity parameters such that \( 0 \leq \epsilon_1 < \epsilon_2 \leq 1 \).

In the following, we formally prove results regarding the above two problems in deterministic and randomized settings, as stated in Section 2.4.

6.1 Deterministic Communication Complexity of GI

**Theorem 6.1.** Deterministic communication complexity of both the tolerant graph isomorphism and non-tolerant graph isomorphism is \( \Theta(n^2) \).

**Proof.** Note that the upper bound protocol (with \( O(n^2) \) bits of communication) is trivial for the problems. So, we only prove the lower bound of \( \Omega(n^2) \) on the communication complexity of the problems. Since the non-tolerant graph isomorphism problem is a special case of the tolerant...
graph isomorphism Problem, it is enough to show that the lower bound for the non-tolerant graph isomorphism Problem.

Note that the non-tolerant graph isomorphism problem is basically checking if the GI-distance between the graphs of Alice and Bob is less than $cn^2$. We will prove the lower bound using a reduction. In the work of Ambainis, Gasarch, Srinavasan and Utis [AGSU15], the authors showed that the Hamming distance problem - where Alice and Bob has two strings of length $m$ and they want to check if the Hamming distance between the strings is less than $A$, for a given $1 \leq A \leq m - 1$ - has communication complexity $\Omega(m)$. We will show that the communication problem of checking if the two graphs have GI-distance less than $A$ is as hard as testing if the Hamming distance between two strings is less than $A$.

Let Alice and Bob have two strings $X$ and $Y$ respectively, with $|X| = |Y| = \binom{n}{2}$. So, we can think of $X$ and $Y$ to represent two graph $G_X$ and $G_Y$ respectively on vertex set of size $n$. Let us construct two graph $G_a$ and $G_b$ as follows:

- $G_a$ and $G_b$ are two graphs on vertex set of size $100n$. Let the vertices be $\{v_1, \ldots, v_{100n}\}$.
- The induced graphs on the first $n$ vertices is $G_X$ and $G_Y$ respectively, That is, $G_a$ and $G_b$ restricted to vertices $\{v_1, \ldots, v_n\}$ is $G_X$ and $G_Y$ respectively.
- For all other pairs of vertices (that is, $(v_i, v_j)$ where at least one of $i$ and $j$ is greater than $n$), we flip an unbiased coin. If the coin turns HEAD, we put the edge $(v_i, v_j)$ in $G_a$ and also in $G_b$, and if the the coin turns TAIL, we do not put the edge $(v_i, v_j)$ in $G_a$ and also not in $G_b$.

So, $G_a$ and $G_b$ are probabilistic construction of two graphs. Using simple probabilistic method technique, we will now show that there exists graphs $G_a$ and $G_b$ such that $d(G_a, G_b) \geq d_H(X, Y)$, where $d_H(X, Y)$ is the Hamming distance between the strings $X$ and $Y$. This would conclude the reduction and the lower bound would follow.

Recall from Definition 2.2 the distance between two graphs under a particular permutation $\sigma$ on the vertices of the graph. If $G_a$ and $G_b$ are two graphs and $P$ be a set of pairs of vertices, then we extend the notation to $d_\sigma(G_a, G_b)|P$ to denote the number of mismatches in entries corresponding to $P$ of the adjacency matrices of $\sigma(G_a)$ and $G_b$. Given a permutation $\sigma$ of the vertices of a graph, let the support of $\sigma$ be defined as the set of vertices moved by the permutation. So the support of $\sigma$ acting on the vertex set $\{v_1, \ldots, v_{100n}\}$ is the set $\{v_i : \sigma(v_i) \neq v_i\}$.

Note that the permutation $\sigma$ induces a permutation on the pair of vertices. Let us denote the permutation to be $\sigma_P$. Let $P_1$ be the set of pairs of vertices where both the vertices are from the set $\{v_1, \ldots, v_n\}$. Let $P_2 = P \setminus P_1$, where $P$ is the set of all pairs of vertices.

Let $\sigma$ be a permutation of the vertices of $G_a$ with support size $k$. Some useful observations (which we believe follows easily and need no explanation) are presented as follows:

- The number of elements in $P_1$ that are moved by $\sigma_P$ is at most $kn$.
- The number of elements $(v_i, v_j) \in P_2$ such that at least one of $\sigma(v_i)$ and $\sigma(v_j)$ is not in $\{v_1, \ldots, v_n\}$ is at least $99nk$.
- There are at least $99nk/3 = 33nk$ pair of disjoint edges from the set $P_2$ such that the first element of the pair is mapped to the second element of the pair by $\sigma_P$.
- Now, since the edges in $P_2$ were put uniformly and independently at random, if $(v_i, v_j)$ and $(\sigma(v_i), \sigma(v_j))$ is one such pair (as described in the previous bullet) then probability that $(v_i, v_j) \in E(G_a)$ iff $(v_i, v_j) \in E(G_b)$ is 1/2. So, we have

$$\mathbb{P}(d_\sigma(G_a, G_b) \leq kn) \leq \left(\frac{33nk}{nk}\right) \frac{1}{2^{233nk-nk}}$$  \hspace{1cm} (15)
From Equation (15) and using union bound, we can say that
\[ P(\forall \sigma, \text{with support size } k, d_\phi(G_a, G_b) \mid P \leq kn) \leq \frac{1}{2^{25nk}} \]  
(16)

Now note that, since we know that an \( \sigma \) with support size \( k \) touches at most \( kn \) elements in \( P_1 \) and \( P_1 \) corresponds to the indices of the strings \( X \) and \( Y \), so from Equation (16) we have that
\[ P(\forall \sigma, \text{with support size } k, d_\phi(G_a, G_b) \leq d_H(X, Y)) \leq \frac{1}{2^{25nk}} \]
(17)

By summing up over all possible values of \( k \in [1, 100n] \), we can say that
\[ P(\forall \sigma, d_\phi(G_a, G_b) \leq d_H(X, Y)) < 1 \]
(18)

Thus, there exists a pair of graphs \( G_a \) and \( G_b \) (as constructed earlier) such that for any permutation of the vertices, the distance increases. So, \( d(G_a, G_b) \geq d_H(X, Y) \). This concludes the proof.

6.2 Randomized Communication Complexity of Non-tolerant GI

Once again we study the non-tolerant graph isomorphism problem. However, unlike above, Alice and Bob can now determine their messages not only from their inputs and previous messages but also with the help of a shared random string.

**Theorem 6.2.** Randomized communication complexity of non-tolerant graph isomorphism is \( \Theta(1) \) in shared randomness model.

**Proof.** Given the graph \( G_a \), Alice will first apply all \( n! \) permutations on the vertices of \( G_a \) and keep this collection of permuted graphs in a set \( X_{G_a} \). She will then represent each graph of this collection as a string of \( O(n^2) \) length with \( \{0, 1\} \) entries. Once she has this collection of permuted copies of \( G_a \) represented as a collection of string over \( \{0, 1\} \), she defines a lexicographic ordering over these strings of \( X_{G_a} \).

Similarly, Bob also performs the steps mentioned above and obtains a collection of strings \( Y_{G_b} \) of the permuted copies of his graph \( G_b \) ordered in lexicographic manner.

Now, they check whether the first strings of the collection of graphs \( X_{G_a} \) and \( Y_{G_b} \) are equal or not. If they are equal, they decide that \( G_a \) and \( G_b \) are isomorphic. Otherwise, they say that \( G_a \) and \( G_b \) are \( \epsilon \) far from isomorphic.

Note that equality of two strings can be decided by \( \theta(1) \) bits of communication in shared randomness model. For correctness of this protocol, note that if the two graphs \( G_a \) and \( G_b \) are isomorphic, the first strings in the lexicographic ordered sets \( X_{G_a} \) and \( Y_{G_b} \) will be equal. Since we do not care about the case when the distance is in between 0 and \( \epsilon n^2 \), we can safely say that \( G_a \) and \( G_b \) are \( \epsilon \) far when the two strings are not equal.

Note that, in the above lemma, we considered the case that the two players Alice and Bob are using shared random bits. However, if they use private random bits, from the seminal result of Newman [New91], we can say that the total amount of communication can only be increased by \( O(\log n) \) bits.
7 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) tolerant 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 2.13, 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 \(\Omega\left(\frac{n}{\log n}\right)\). 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?**

As mentioned before, it is interesting to note that our lower bound proof is via a pure reduction from tolerant graph isomorphism to tolerant testing of EMD of multi-sets over the Hamming cube using samples without replacement. Using our lower bound technique (and Proposition 2.10), 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 tolerant testing of 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 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, and

(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 and communication world. So, another natural open question to look for is:

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

References


