Interplay Between Graph Isomorphism and Earth Mover's Distance in the Query and Communication Worlds

Sourav Chakraborty
Indian Statistical Institute, Kolkata, India

Arijit Ghosh
Indian Statistical Institute, Kolkata, India

Gopinath Mishra
Indian Statistical Institute, Kolkata, India

Sayantan Sen
Indian Statistical Institute, Kolkata, India

Abstract

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

In this paper, we 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.

2012 ACM Subject Classification Theory of computation → Streaming, sublinear and near linear time algorithms

Keywords and phrases Graph Isomorphism, Earth Mover Distance, Query Complexity

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2021.34

Category RANDOM

Related Version *Full Version*: https://eccc.weizmann.ac.il/report/2020/135/

Acknowledgements The authors would like to thank an anonymous reviewer for pointing out a mistake in an earlier version of this paper, as well as the reviewers of RANDOM for various suggestions that improved the presentation of the paper.

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2021).
Editors: Mary Wootters and Laura Sanità; Article No. 34; pp. 34:1–34:23
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Graph isomorphism (GI) has been one of the most celebrated problems in computer science. Roughly speaking, the graph isomorphism problem asks whether two graphs are structure-preserving. Namely, given two graphs G_u and G_k, graph isomorphism of G_u and G_k is a bijection $\psi : V(G_u) \to V(G_k)$ such that for all pair of vertices $u, v \in V(G_u)$, the edges $\{u, v\} \in E(G_u)$ if and only if $\{\psi(u), \psi(v)\} \in E(G_k)$ [1]. One central open problem in complexity theory is whether the graph isomorphism problem can be solved in polynomial time. Recently in a breakthrough result, Babai [5] 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 [7] and PhD thesis of Sun [24]).

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

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

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

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

The problem of computing GI-distance between two graphs is known to be $\#P$-hard [18]. 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 [18], 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 and two party communication complexity model.

1.1 Property Testing of Graph Isomorphism

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

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

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

In the query model, the problem is equivalent (up to a constant factor) to the tolerant property testing of graph isomorphism in the dense graph model (introduced in the work of Parnas, Ron and Rubinfeld [21]). For $0 \leq \gamma < 1$, two graphs G_u and G_k, with n vertices, are called γ-*close* or γ-*far to isomorphic*[2] if $d(G_u, G_k) \leq \gamma n^2$ or $d(G_u, G_k) \geq \gamma n^2$, respectively. In (γ_1, γ_2)-*tolerant GI testing*, we are given two graphs G_u and G_k, and two parameters $0 \leq \gamma_1 < \gamma_2 \leq 1$, with the guarantee that either the graphs are γ_1-close or γ_2-far. One of the graphs (usually denoted as G_u) is accessed by querying the entries of its adjacency matrix. In contrast, the other graph (usually denoted as G_k [3]) is known to the query algorithm, and no cost for accessing the entries of the adjacency matrix of G_k is incurred. The query complexity is the number of queries (to the adjacency matrix of G_u) that are required for testing, (with correctness probability at least 2/3 [4]), whether G_u and G_k are γ_1-close or γ_2-far. The query algorithm is assumed to have unbounded computational power.

The non-tolerant property testing version of the graph isomorphism problem (that is, when $\gamma_1 = 0$) was first studied by Fischer and Matsliah [13] and subsequently, Babai and Chakraborty [6] 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 [15] studied the problem for the *bounded degree graph model* of property testing and Levi and Medina [17] 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 [13] (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 [13, 19]. 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 $\widetilde{\mathcal{O}}(n)$ for tolerant testing of graph isomorphism (when one of the graphs is known in advance). Basically, one goes over all possible $n!$ bijections $\phi : V_u \to V_k$ 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 $\widetilde{\mathcal{O}}(n)$ queries and outputs a number α such that, with probability at least 2/3, the following holds:

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

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).

---

[2] As a shorthand, rather than saying γ-close or γ-far to isomorphic, we will just say γ-*close* or γ-*far* respectively.
[3] G_u and G_k denote the unknown and known graphs, respectively.
[4] The correctness probability can be made any $1 - \delta$ by incurring a multiplicative factor of $O(\log \frac{1}{\delta})$ in the query complexity.
[5] If the samples are $\Theta(\log(n!))$, then the error probability can be bounded using the union bound.

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 [13] 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 $\widetilde{\mathcal{O}}(q(n))$ queries, where the tilde hides a polylogarithmic factor of n (number of vertices). And since the query complexity for testing identity of distributions (from [8], [20], [1], [26]) is known to be $\mathcal{O}(\frac{\sqrt{n}}{\epsilon^2})$, the query complexity for non tolerant GI-testing is $\widetilde{\mathcal{O}}(\sqrt{n})$.

In the lower bound proof of [13], 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.

1.2 Earth Mover's Distance (EMD)

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

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

A standard way to think of sampling from any probability distribution is to consider it as a 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 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 β_1, β_2 such that $0 \le \beta_1 < \beta_2 \le 1$, a known multi-set S_k and access to the unknown multi-set S_u by sampling **with** (or **without**) replacement, decides whether $EMD(S_k, S_u) \le \beta_1 n^2$ or $EMD(S_k, S_u) \ge \beta_2 n^2$ 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.

[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 ϵ.

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

Earth Mover's Distance (EMD) is a fundamental metric over the space of distributions supported on a fixed metric space. Estimating EMD between two distributions, up to a multiplicative factor, has been extensively studied in mathematics and computer science. It is closely related to the embedding of the EMD metric into a ℓ_1 metric. Even the problem of estimation of EMD between distributions up to an additive factor has been well studied, for reference see [12], [23]. The hardness of estimating EMD between distributions depends heavily on the structure of the domain on which the distributions are supported. In [12], the authors have proved a lower bound of $\Omega((\Delta/\epsilon)^d)$ on the query complexity for estimating (up to an additive error of ϵ) EMD between two distributions supported on the real cube $[0, \Delta]^d$. At the same time, it is not hard to see that if the support has certain structures, estimating EMD may be easy. In this paper, we focus on the estimation of EMD between two distribution when the metric space is the Hamming cube.

As noted earlier, sample access to a probability distribution is precisely the same as uniform sampling from a multi-set **with** replacement. Thus, from the results of Valiant and Valiant [25], it can be shown that the sample complexity for estimating the EMD between two distribution over the Hamming cube of dimension n is $\Omega(n/\log n)$. In other words, $\mathrm{QWR}_{\mathrm{EMD}}(n) = \Omega(n/\log n)$, and this is tight ignoring polynomial factor in $\log n$ (See Theorem B.10 of Appendix B). But what about $\mathrm{QWoR}_{\mathrm{EMD}}(n)$? To the best of our knowledge, the sample complexity measure when the distributions are accessed by sampling a multi-set **without** replacement has never been studied before (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 [22] for proving a lower bound of distinct elements problem. Also, recently Goldreich [15] considered a similar sampling **without** replacement model while studying the non-tolerant graph isomorphism in the bounded degree model.

Coming back to our context, it can be proven that: if $\mathrm{QWoR}_{\mathrm{EMD}}(n) = o(\sqrt{n})$, then $\mathrm{QWR}_{\mathrm{EMD}}(n) = o(\sqrt{n})$ (See Proposition B.7 of Appendix B). As $\mathrm{QWR}_{\mathrm{EMD}}(n) = \Omega(\frac{n}{\log n})$, we have a lower bound of $\Omega(\sqrt{n})$ on $\mathrm{QWoR}_{\mathrm{EMD}}(n)$. To the best of our knowledge, there is no known better lower bound than $\Omega(\sqrt{n})$ for $\mathrm{QWoR}_{\mathrm{EMD}}(n)$, although a lower bound of $\Omega(\frac{n}{\log n})$ exists for $\mathrm{QWR}_{\mathrm{EMD}}(n)$ (using observation in [12]). We verified that the proof of [27] also goes through for $\mathrm{QWoR}_{\mathrm{EMD}}(n)$ as well (See Theorem 1.5). We now present the following conjecture:

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

One of our main contributions in this paper is introducing this complexity measure of $\mathrm{QWoR}_{\mathrm{EMD}}(n)$ as well as the above conjecture. In the rest of the paper, we focus on exploring the connection between $\mathrm{QWoR}_{\mathrm{EMD}}(n)$ and the query complexity of tolerant GI-testing. For a formal discussion on EMD over Hamming cube, please refer to Appendix B.

1.3 Our Results

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

▶ **Theorem 1.4** (Main Result). *Let G_k and G_u denote the known and the unknown graphs on n vertices, respectively, and $Q_{GI}(G_u, G_k)$ denotes the number of adjacency queries to G_u, required by the best algorithm that takes two constants γ_1, γ_2 with $0 \leq \gamma_1 < \gamma_2 \leq 1$ and decides whether $d(G_u, G_k) \leq \gamma_1 n^2$ or $d(G_u, G_k) \geq \gamma_2 n^2$ with probability at least $2/3$. Then*

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

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

1.3.1 Implication of Theorem 1.4 to Query Complexity of Tolerant GI

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. Thus our reductions also hold for other computational models such as the communication complexity model. Regarding the lower bound on the sample complexity of tolerant EMD testing of multi-sets (in the **with** replacement model), using observation in [12], we note that the tolerant EMD testing is as hard as tolerant testing of variation distance. In [27], they gave a lower bound of $\Omega(n^{1-o(1)})$ on the sample complexity for tolerant ℓ_1 testing. Although the proof of [27] uses samples **with** replacement (when we think of a distribution as a multi-set), it can be verified that the proof also works for samples **without** replacement.

▶ **Theorem 1.5** (Follows from [27]). *For any constants $0 < \alpha < \beta < 1$, distinguishing between distribution pairs with statistical distance less than α from those with distance greater than β requires $n^{1-o(1)}$ samples* **without** *replacement.*

From Theorem 1.5, a similar lower bound follows for tolerant EMD testing of multi-sets **without** replacement. Thus, from Theorem 1.4, we have the following corollary:

▶ **Corollary 1.6.** *Let G_k and G_u be the known and unknown graphs on n vertices, respectively. For any constants $0 < \gamma_1 < \gamma_2 < 1$, distinguishing between isomorphism distance of $d(G_u, G_k) \leq \gamma_1 n^2$ with $d(G_u, G_k) \geq \gamma_2 n^2$ requires $n^{1-o(1)}$ queries to the adjacency matrix of G_u. On the other hand, for any constants $0 < \gamma_1 < \gamma_2 < 1$, distinguishing between isomorphism distance of $d(G_u, G_k) \leq \gamma_1 n^2$ with $d(G_u, G_k) \geq \gamma_2 n^2$ can be done in $\widetilde{O}(n)$ queries.*

The lower bound of [27] was later improved to $\Omega(\frac{n}{\log n})$ in [25]. However, the arguments of [25] are much more delicate and it is not completely clear to us whether their result of $\Omega(\frac{n}{\log n})$ can be carried over to the **without** replacement setting, even if we allow a loss of polylogarithmic factor. So, we propose the following conjecture:

▶ **Conjecture 2.** *Let G_k and G_u be the known and unknown graphs on n vertices, respectively. For any constants $0 < \gamma_1 < \gamma_2 < 1$, distinguishing between isomorphism distance of $d(G_u, G_k) \leq \gamma_1 n^2$ with $d(G_u, G_k) \geq \gamma_2 n^2$ requires $\Omega(\frac{n}{\log n})$ queries to the adjacency matrix of G_u.*

Note that Conjecture 1 and Conjecture 2 are equivalent. Besides, the difference between sampling **with** and **without** replacement is much more subtle. Freedman [14] 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 ℓ_1 distance) to the distribution of answers to the queries when samples are drawn **without** replacement. Thus, following Proposition B.7 along with Theorem 1.4, we can get an alternative proof of the following lower bound proved by Fischer and Matsliah [13].

▶ **Corollary 1.7** (Fischer and Matsliah [13]). *There exists a constant $\zeta \in (0,1)$ such that any query algorithm that decides, with probability at least $2/3$, if a known graph G_k and an unknown graph G_u is isomorphic or γ-far from isomorphic, with $\gamma \leq \zeta$, must make $\Omega(\sqrt{n})$ queries.*

1.3.2 Implication of Theorem 1.4 to Communication Complexity of Tolerant GI

One of the central models of computation (particularly in the context of theoretical computer science) is the 2-player communication game introduced by Yao [28] in 1979. 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 $O(1)$ (see the full version for the proof).

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 $d(G_a, G_b) \geq \gamma_2 n^2$? While we don't have a complete answer to this question yet, the following theorem holds from Theorem 1.2:

▶ **Theorem 1.8** (Informally stated). *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 γ_1-close or γ_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 [7] 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 $\widetilde{\Theta}\left(c(n, \gamma_1', \gamma_2')\right)$, where γ_1' and γ_2' are constants that depend only on β_1 and β_2, and $\widetilde{\Theta}(\cdot)$ hides multiplicative factor of* $\mathrm{poly}(\log n)$.

Theorem 1.8 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 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 ([27] and [25]) does not give a lower bound

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

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 [16, 3, 2, 4]. 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 1.9.** *Deterministic communication complexity of non-tolerant version of Graph Isomorphism testing (hence the tolerant version) is* $\Theta(n^2)$.

The proof of the above theorem is present in the full version of the paper [10].

Organization of the paper. In Section 2, 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 3 and 4 respectively. We finally conclude in Section 5. For space constraint, we could not add all possible proofs. Please see [10] for the full version of the paper.

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

2 Discussion on our proof of Theorem 1.4

2.1 Reduction from tolerant EMD testing to tolerant graph isomorphism testing (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 3.3) 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 π is applied to the vertices of G_u.

The permutation π 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 π, 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 π 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 Φ 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 $d_\Phi(G_k, G_u) = 2 \cdot EMD(S_k, S_u)$, where $d_\Phi(G_k, G_u) = \min_{\phi \in \Phi} d_\phi(G_k, G_u)$ (See [10], Lemma 3.5 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 [10] Lemma 3.6 for a formal proof).

One might compare our proof technique to the lower bound proof of (non-tolerant) testing of GI from [13]. In [13], $\Omega(\sqrt{n})$ 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 **without** replacement. This can be of much importance, mainly while considering other models of computation, like in the communication model. From our reduction, we can obtain an alternative proof of $\Omega(\sqrt{n})$ lower bound for the (non-tolerant) GI testing via the $\Omega(\sqrt{n})$ lower bound of the equivalence testing of distributions, as pointed out in Corollary 1.7.

2.2 Reduction from tolerant graph isomorphism to tolerant EMD testing (Upper bound part of Theorem 1.4)

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 1.4.

Algorithm for tolerant graph isomorphism using algorithm for tolerant EMD testing as a black box:

Our testing algorithm is inspired by the algorithm of Fischer and Matsliah [13] 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 $\mathcal{O}\left(\frac{1}{\gamma_2-\gamma_1}\right)$ size collection of random subset of vertices, i.e, *coresets* $\mathcal{C}_u$ from the unknown graph G_u where each $C_u \in \mathcal{C}_u$ is of size $\mathcal{O}(\log n)$. 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 $C_k^i = \eta_i(C_u)$. Now each C_u (as well as each C_k^i) defines a label distribution of the vertices of G_u (as well as G_k). Let us denote the set of labels as X_{C_u} (and $Y_{C_k^i}$). Now we test if the EMD between X_{C_u} and $Y_{C_k^i}$ is close or far for each $i \in [J]$ (See Claim 4.2). We keep only those (C_u, η_i) for Phase 2 such that $EMD(X_{C_u}, Y_{C_k^i}) \leq \left(\gamma_1 + \frac{\gamma_2-\gamma_1}{2000}\right) n\,|C_u|$.

Although Phase 1 of our algorithm is similar to the algorithm of [13], there is a striking difference. Since the authors of [13] were testing the non-tolerant version of graph isomorphism, they were testing the identity of the label distributions of X_{C_u} and $Y_{C_k^i}$. However, since we are solving the tolerant version of the problem, we need to allow some error among the label distributions. We need to pass only those placements of C_u that under *good bijections* do not produce much error and testing of tolerant EMD fits exactly for this purpose. 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 $\mathcal{O}\left(\frac{\log^2 n}{(\gamma_2-\gamma_1)^3}\right)$ many vertices from the unknown graph G_u randomly and call it W. We further find the labels of all the vertices of W under C_u-labelling by querying the corresponding entries of G_u for each C_u that has passed Phase 1. Then we try to match the vertices of W to the set of all possible labels $\{l_1, l_2, \ldots, l_t\}$ of the vertices of G_k under C_k^i-labelling where $C_k^i = \eta_i(C_u)$, for those η_i that have passed Phase 1. Ideally, we would like to find a mapping $\psi : W \to \{l_1, l_2, \ldots, l_t\}$ such that the total distance between the labels of the matched vertices is not too large. If no such ψ is possible, we reject the current embedding and try some other embedding that has passed Phase 1.

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

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. The proofs are present in the full version of the paper [10].

3 Tolerant graph isomorphism is as hard as tolerant EMD testing

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

▶ **Theorem 3.1** (Restatement of the lower bound part of Theorem 1.4). *Let G_k be the known and G_u be the unknown graph on n vertices, where $n \in N$ is sufficiently* large. *There exists a constant $\epsilon_{ISO} \in (0,1)$ such that for any given constants γ_1, γ_2 with $0 < \gamma_1 < \gamma_2 < \epsilon_{ISO}$, any algorithm that decides whether the graphs are γ_1-close or γ_2-far, requires $\mathrm{QWoR}_{\mathrm{EMD}}(n)$ adjacency queries to the unknown graph G_u where $\mathrm{QWoR}_{\mathrm{EMD}}$ is as defined in Definition 1.3.*

In Section 2.1, we have discussed an overview of of our idea to prove the above theorem. To prove Theorem 3.1, we show a reduction from tolerant GI testing to tolerant EMD testing over multi-sets when we have samples **without** replacement from the unknown multi-set.

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

Then, for any constants β_1 and β_2 with $0 < \beta_1 < \beta_2 < \frac{\epsilon_0}{2}$, the following holds where $\kappa = \frac{\beta_2 - \beta_1}{8}$ and $T_\kappa = \lceil \frac{30}{\kappa(2-\kappa)} \rceil$. There is a tolerant tester for EMD such that, given a known and an unknown multi-set S_k and S_u respectively, of the Hamming cube $\{0,1\}^{T_\kappa n}$ with $|S_k| = |S_u| = n$, can distinguish whether $EMD(S_k, S_u) \leq \beta_1 T_\kappa n^2$ or $EMD(S_k, S_u) \geq \beta_2 T_\kappa n^2$ with Q many samples ***without*** *replacement from S_u.*

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

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

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

▶ **Lemma 3.3.** *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 $n \in \mathbb{N}$ [8], 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 $sn - 2$.*

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

The proof of Lemma 3.3 uses probabilistic method (See [10] for the proof). Let $ALG(\gamma_1, \gamma_2, T)$ be the algorithm that takes γ_1 and γ_2 with $0 < \gamma_1 < \gamma_2 < \epsilon_0$ as input and decides whether $d(G_k, G_u) \leq \gamma_1 T n^2$ or $d(G_k, G_u) \geq \gamma_2 T n^2$, where $|V(G_k)| = |V(G_u)| = (T+1)n$. Now we show that for any two constants β_1 and β_2 with $0 < \beta_1 < \beta_2 < \frac{\epsilon_0}{2}$, $\kappa = \frac{\beta_2 - \beta_1}{8}$ and $T_\kappa = \lceil \frac{6s}{\kappa(2-\kappa)} \rceil$, there exists an algorithm $\mathcal{A}(\beta_1, \beta_2, \kappa, T_\kappa)$ that can test whether two multi-sets S_k and S_u over the $T_\kappa n$-dimensional Hamming cube have EMD less than $T_\kappa \beta_1 n^2$ or more than $T_\kappa \beta_2 n^2$ with Q many queries to the multi-set S_u. To be specific, algorithm $\mathcal{A}(\beta_1, \beta_2, \kappa, T_\kappa)$ for EMD testing will use algorithm $ALG(\gamma_1, \gamma_2, T)$ for (γ_1, γ_2)-tolerant GI such that $\gamma_1 = 2\beta_1$, $\gamma_2 = 2\beta_2 - 2\kappa$ and $T = T_\kappa$. Note that, as $0 < \beta_1 < \beta_2 < \frac{\epsilon_0}{2}$ and $\kappa = \frac{\beta_2 - \beta_1}{8}$, $0 < \gamma_1 < \gamma_2 < \epsilon_0$ holds. The details of the reduction, that is, algorithm $\mathcal{A}$ is described below. Because of space constraint, we are not presenting the proof of correctness of the reduction in this extended abstract. Please refer to our full version [10].

Description of the reduction

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

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

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

- $G_k[A_k]$ is a clique with n vertices.
- $G_k[B_k]$ is a copy of the graph $G_p(V_p, E_p)$ on $T_\kappa n$ vertices as stated in Lemma 3.3 with parameters $s = 5$, $\kappa = \frac{\beta_2 - \beta_1}{8}$ 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 3.3, with parameters $s = 5$, $\kappa = \frac{\beta_2 - \beta_1}{8}$ and $T_\kappa = C_{\kappa,5}$.
- Let us first pick a random permutation π on $[n]$. For the cross edges between the vertices in A_u and B_u, we add the edge $(a'_{\pi(i)}, b_j)$ to $E(G_u)$ if and only if the j-th coordinate of u_i is 1 for all $i \in [n]$ and $j \in [T_\kappa n]$.

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

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

Simulating Queries to G_u by samples drawn from S_u *without* replacement

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

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

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

4 Tolerant EMD testing is as hard as tolerant graph isomorphism testing

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

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

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

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

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

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

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

4.1 Algorithm for tolerant graph isomorphism testing

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

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

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

Now we present our query algorithm **TolerantGI($\boldsymbol{G_u}$, $\boldsymbol{G_k}$, $\boldsymbol{\gamma_1}$, $\boldsymbol{\gamma_2}$)** that comprises three phases. The technical overview of the algorithm is already presented in Section 2.2

Formal Description of TolerantGI($\boldsymbol{G_u}$, $\boldsymbol{G_k}$, $\boldsymbol{\gamma_1}$, $\boldsymbol{\gamma_2}$):

The three phases of our algorithm are as follows:

4.1.1 Phase 1

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

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

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

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

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

4.1.2 Phase 2

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

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

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

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

$$\sum_{w\in W} d_H(\mathcal{L}_{C_u}(w), \psi(w)) \le \frac{2\gamma_2}{5}\,|C_u|\,|W| \text{ and } |\{w : \psi(w) = l_j\}| \le |V_j|\,\forall\, j \in [t]. \quad (1)$$

Let Γ_W be the set of tuples such that

$$\Gamma_W = \{(C_u, \eta_i, \psi) : (C_u, \eta_i) \in \Gamma \text{ and } \psi \text{ satisfies Equation (1)}\}.$$

4.1.3 Phase 3

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

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

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

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

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

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

5 Conclusion

In this paper, we proved that the query complexity of tolerant GI testing between a known graph G_k and an unknown graph G_u is the same as (up to polylogarithmic factor) tolerant testing of EMD between a known multi-set S_k and an unknown multi-set S_u when we have

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

samples **without** replacement from S_u. In Lemma B.10, we have shown that the sample complexity of testing of EMD between a known multi-set S_k and an unknown multi-set S_u when we have samples **with** replacement from S_u is $\Omega(n/\log n)$. Thus the natural open question is

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

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 B.7), we can get an alternative proof of Fischer and Matsliah's lower bound result for testing non-tolerant graph isomorphism [13]. 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 [13] studied the non-tolerant version of the graph isomorphism problem in two scenarios: (i) one graph is known and the other graph is unknown, (ii) both the graphs are unknown. They resolved the query complexity of (i), whereas Onak and Sun [19] 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

1 Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. *arXiv preprint arXiv:1507.05952*, 2015.

2 Alexandr Andoni, Khanh Do Ba, Piotr Indyk, and David Woodruff. Efficient sketches for earth-mover distance, with applications. In *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 324–330. IEEE, 2009.

3 Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Earth mover distance over high-dimensional spaces. In *SODA*, volume 8, pages 343–352, 2008.

4 Alexandr Andoni, Robert Krauthgamer, and Ilya Razenshteyn. Sketching and embedding are equivalent for norms. *SIAM Journal on Computing*, 47(3):890–916, 2018.

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

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

7 László Babai, Anuj Dawar, Pascal Schweitzer, and Jacobo Torán. The Graph Isomorphism Problem (Dagstuhl Seminar 15511). *Dagstuhl Reports*, 5(12):1–17, 2015. doi:10.4230/DagRep.5.12.1.

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

9 Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020.

10 Sourav Chakraborty, Arijit Ghosh, Gopinath Mishra, and Sayantan Sen. Interplay between graph isomorphism and earth mover's distance in the query and communication worlds. In *Electron. Colloquium Comput. Complex.*, volume 27, page 135, 2020.

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

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

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

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

15 Oded Goldreich. Testing isomorphism in the bounded-degree graph model. *Electron. Colloquium Comput. Complex.*, 26:102, 2019. URL: https://eccc.weizmann.ac.il/report/2019/102.

16 Subhash Khot and Assaf Naor. Nonembeddability theorems via fourier analysis. *Mathematische Annalen*, 334(4):821–852, 2006.

17 Reut Levi and Moti Medina. Distributed testing of graph isomorphism in the congest model. *arXiv preprint arXiv:2003.00468*, 2020.

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

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

20 Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

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

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

23 Shashank Singh and Barnabás Póczos. Minimax distribution estimation in wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

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

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

26 Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

27 Paul Valiant. Testing Symmetric Properties of Distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.

28 Andrew Chi-Chih Yao. Some complexity questions related to distributive computing (preliminary report). In Michael J. Fischer, Richard A. DeMillo, Nancy A. Lynch, Walter A. Burkhard, and Alfred V. Aho, editors, *Proceedings of the 11h Annual ACM Symposium on*

Theory of Computing, April 30 - May 2, 1979, Atlanta, Georgia, USA, pages 209–213. ACM, 1979. doi:10.1145/800135.804414.

A Preliminaries

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

A.1 Notion of distance between two graphs

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

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

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

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

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

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

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

Next we define the concept of closeness between two graphs.

▶ **Definition A.3.** (CLOSE and FAR) For $\gamma \in [0, 1)$, two graphs G_u and G_k with n vertices are *γ-close to isomorphic* if $d(G_u, G_k) \le \gamma n^2$. Otherwise, we say G_u and G_k are *γ-far from being isomorphic.* [10]

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

A.2 Property Testing of Distribution Properties

Understanding different properties of probability distributions have been an active area of research in property testing (For reference, see [9]). 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 [14], we know that the variation distance between probability distributions when accessed via samples **with** and **without** replacement, becomes arbitrary close to $1/2$ when the number of samples is $\Omega(\sqrt{n})$. Because of this reason, many techniques developed for sampling **with** replacement for various problems no longer work anymore. Most importantly, proving any lower bound better than $\Omega(\sqrt{n})$ is often nontrivial.

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

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

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

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

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

$$\text{Minimize} \quad \sum_{x,y \in H} f_{xy} d_H(x,y)$$

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

Now we define EMD between two multi-sets.

▶ **Definition B.3** (EMD between two multi-sets)**.** Let S_1, S_2 be two multi-sets on a Hamming cube $H = \{0,1\}^d$ of dimension d with $|S_1| = |S_2|$. The EMD between S_1 and S_2 is denoted by $EMD(S_1, S_2)$ and defined as $EMD(S_1, S_2) = \min_{\phi: S_1 \to S_2} \sum_{x \in S_1} d_H(x, \phi(x))$ where ϕ 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 B.4** (Query accesses to multi-sets)**.** A multi-set S of n elements is accessed in one of the following ways:

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

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

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

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

For a formal proof of the above proposition, see [10]. The following observation connects the EMD between two probability distributions with that of between two multi-sets.

▶ **Observation B.6.** Let p, q be two K-grained probability distributions [11] 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, $EMD(p,q) = \frac{EMD(S_1,S_2)}{K}$.

See [10] for a formal proof.

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

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

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

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

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

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

The problem of estimating the EMD over multi-sets while sampling **with** (or **without**) replacement means designing an algorithm, that given any two constants β_1, β_2 such that $0 \leq \beta_1 < \beta_2 \leq 1$, and access to the unknown set S_u by sampling **with** (or **without**)

[11] The probability of each element in the sample space is an integer multiple of $\frac{1}{K}$.

replacement decides whether $EMD(S_k, S_u) \leq \beta_1 n^2$ or $EMD(S_k, S_u) \geq \beta_2 n^2$ with probability at least $2/3$.

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

Let $\text{QWR}_{\text{EMD}}(n, d, \beta_1, \beta_2)$ (and $\text{QWoR}_{\text{EMD}}(n, d, \beta_1, \beta_2)$) 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 $\text{QWoR}_{\text{EMD}}(n, d)$ ($\text{QWR}_{\text{EMD}}(n, d)$) instead of $\text{QWoR}_{\text{EMD}}(n, d)$ ($\text{QWR}_{\text{EMD}}(n, \beta_1, \beta_2)$) when the proximity parameters are clear from the context.

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

(i) $\text{QWR}_{\text{EMD}}(n_1, d) \leq \text{QWR}_{\text{EMD}}(n_2, d)$*;*
(ii) $\text{QWoR}_{\text{EMD}}(n_1, d) \leq \text{QWoR}_{\text{EMD}}(n_2, d)$*;*
(iii) $\text{QWR}_{\text{EMD}}(n, d_1) \leq \text{QWR}_{\text{EMD}}(n, d_2)$*; and*
(iv) $\text{QWoR}_{\text{EMD}}(n, d_1) \leq \text{QWoR}_{\text{EMD}}(n, d_2)$*.*

▶ **Remark 7.** For $d = n$ (as considered in Definition 1.3), $\text{QWoR}_{\text{EMD}}(n, d)$ (and $\text{QWR}_{\text{EMD}}(n, d)$) are denoted as $\text{QWoR}_{\text{EMD}}(n)$ (and $\text{QWR}_{\text{EMD}}(n)$).

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

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

Thus following Proposition B.7, we have

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

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

▶ **Lemma B.12.** *Let 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 ϵ_{EMD} such that the following holds. Given two constants β_1, β_2 with $0 < \beta_1 < \beta_2 < \epsilon_{EMD}(c)$, $\Omega\left(\frac{n}{\log n}\right)$ samples from the distribution q are necessary in order to decide whether $EMD(p, q) \leq \beta_1 n$ or $EMD(p, q) \geq \beta_2 n$. More over, $\epsilon_{EMD} = \frac{1-\epsilon_{\ell_1}}{4}$, where ϵ_{ℓ_1} is the constant that is mentioned in Theorem B.14.*

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

▶ **Theorem B.13** (Valiant and Valiant [25])**.** *Let p and q be two known and unknown probability distributions respectively over $[n]$. There is an absolute constant ϵ such that in order to decide whether $\|p - q\|_1 \leq \epsilon$ or $\|p - q\|_1 \geq 1 - \epsilon$, $\Omega(\frac{n}{\log n})$ samples, from the distribution q, are necessary.* [12]

[12] Note that this is rephrasing of the result proved in [25]. For reference, see Chapter 5 of the survey by Canonne [9].

Now, we restate the above result for our purpose.

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

As noted earlier, we will prove Theorem B.10 by using Lemma B.14. However, Theorem B.10 is regarding EMD between two distributions whereas Lemma B.14 is regarding ℓ_1 distance between two distributions. The following observation (from [12]) gives a connection between EMD between two distributions with the ℓ_1 distance between them, which will be required in lower bound proof.

▶ **Proposition B.15** ([12]). *Let (M, D) be a finite metric space and p and q be two probability distributions on M. Minimum distance between any two points of M is $\Delta_{\min}$ and diameter of M is $\Delta_{\max}$. Then the following condition holds:*

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

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

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

Algorithm $\mathcal{P}$:

(1) First run algorithm $\mathcal{A}$.
(2) If the output of algorithm $\mathcal{A}$ is $EMD(p,q) \leq \beta_1 n$, algorithm $\mathcal{P}$ returns $\|p-q\|_1 \leq \alpha_1$.
(3) If the output of algorithm $\mathcal{A}$ is $EMD(p,q) \geq \beta_2 n$, algorithm $\mathcal{P}$ returns $\|p-q\|_1 \geq \alpha_2$.

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

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

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

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

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

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

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

Proof. Let ϵ be a constant less than $(\beta_2 - \beta_1)$. We construct a probability distribution q' such that the ℓ_1 distance between q and q' will be at most ϵ, that is, $\sum\limits_{i \in [L]} |q(i) - q'(i)| \leq \epsilon$. Note that such a q' can be constructed with probability at least $1 - \delta$ by querying for $\widetilde{\mathcal{O}}(n)$ many samples of q which follows from [11]. Then, we find $EMD(p,q')$. Observe that $|EMD(p,q) - EMD(p,q')| \leq \frac{\epsilon n}{2}$. This is because

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

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