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

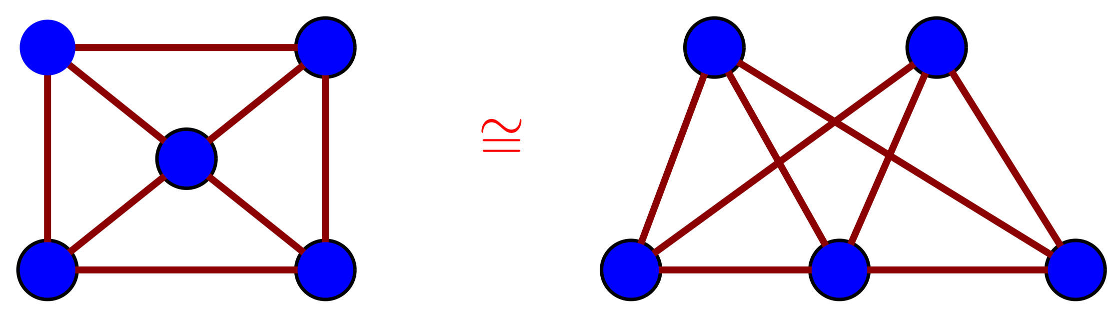
Sourav Chakraborty, Arijit Ghosh, Gopinath Mishra, Sayantan Sen

Indian Statistical Institute, Kolkata, India

Tolerant Testing of Graph Isomorphism

Graph Isomorphism

Graphs G_u and G_k are isomorphic if there exists 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 $\{\psi(u), \psi(v)\} \in E(G_k)$.



GI Distance

For a bijection $\phi : V_u \to V_k$, the distance between G_u and G_k is defined as

$$d_\phi(G_u, G_k) := \left| \left\{ (u, v) : \begin{array}{l} \text{Exactly one among } (u,v) \in E_u \\ \text{or } (\phi(u), \phi(v)) \in E_k \text{ holds} \end{array} \right\} \right|$$

The GI Distance between G_u and G_k is defined as

$$d(G_u, G_k) := \min_{\phi : V_u \to V_k} d_\phi(G_u, G_k)$$

How to access a graph?

- Graphs can be accessed by querying the entries of the adjacency matrix.

(ϵ_1, ϵ_2)-GI Testing

For a known graph G_k and an unknown graph G_u (both on n vertices) and two proximity parameters ϵ_1 and ϵ_2 with $0 \le \epsilon_1 < \epsilon_2 \le 1$, want to decide whether $d(G_u, G_k) \le \epsilon_1 n^2$ or $d(G_u, G_k) \ge \epsilon_2 n^2$?

What is the query complexity of (ϵ_1, ϵ_2)-GI testing?

Known Result

Fischer-Matsliah (SICOMP'08) proved that the query complexity of non-tolerant GI (when $\epsilon_1 = 0$), is $\widetilde{\Theta}(\sqrt{n})$.

Tolerant Testing of EMD

Earth Mover Distance (EMD)

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

$$\textbf{min} \sum_{i,j \in H} f_{ij} d_H(i,j)$$

$$\textbf{Subject to} \sum_{j \in H} f_{ij} = p_u(i) \; \forall i \in H \;\; \& \;\; \sum_{i \in H} f_{ij} = p_k(j) \; \forall j \in H$$

How to access a distribution?

- Accessing a distribution via samples is equivalent to samples **with** replacement from a multi-set. Thus

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

- For multi-sets, both sampling **with** and **without** replacement are possible. We focus on sampling **without** replacement model here.

(γ_1, γ_2)-EMD Testing

For a known probability distribution p_k and an unknown probability distribution p_u, accessed by samples **without** replacement, want to decide whether $EMD(p_u, p_k) \le \gamma_1$ or $EMD(p_u, p_k) \ge \gamma_2$ for $0 \le \gamma_1 < \gamma_2 \le 1$?

What is the minimum number of samples WITHOUT replacement required for (γ_1, γ_2)-EMD testing?

Known Results

- $O(n)$ samples **without** replacement are enough to decide (γ_1, γ_2)-EMD. This follows from learning unknown distribution
- The best lower bound of (γ_1, γ_2)-EMD is $\Omega(n^{1-o(1)})$ samples **without** replacement following Valiant (STOC'08).

Theorem: Tolerant GI Testing ≡ Tolerant EMD Testing WITHOUT replacement

Let $Q_{GI}(G_u, G_k)$ denote the number of adjacency queries to G_u, required to decide (ϵ_1, ϵ_2)-GI and $QWOR_{EMD}$ denote the number of samples **without** replacement required to decide (γ_1, γ_2)-EMD. Then

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

Sketch of Proof

- **Upper bound**: Non trivial generalization of Fischer-Matsliah's algorithm of non-tolerant testing.
- **Lower bound**: Follows from a *pure* reduction.

Implications

- The query complexity for tolerant GI testing is $\Omega(n^{1-o(1)})$.
- Lower bound reduction technique holds for other computation models, say for communication complexity model.
- Using our lower bound reduction, we give a proof of Fischer-Matsliah's lower bound for $(0, \epsilon_2)$-GI.

Communication Complexity

- **Question:** Alice and Bob have graphs G_A and G_B respectively and want to decide if $d(G_A, G_B)$ is less than ϵ_1 or more than ϵ_2 by communicating among themselves.
- Randomized communication complexity of non tolerant GI is $O(1)$ whereas deterministic communication complexity of non tolerant GI (hence tolerant GI) is $\Omega(n^2)$
- Our theorem gives an equivalence of tolerant GI and tolerant EMD in the communication setting.

Open Problems

- Can Valiant-Valiant's lower bound (STOC'11) of $\Omega(\frac{n}{\log n})$ for (γ_1, γ_2)-EMD testing **with** replacement be extended to **without** replacement?
- What is the randomized communication complexity of (ϵ_1, ϵ_2)-GI?