SUPPLEMENTARY MATERIAL
OPTIMAL TRANSPORT-BASED SUPERVISED GRAPH SUMMARIZATION

Anonymous authors
Paper under double-blind review

1 MNIST RUNTIME COMPARISON

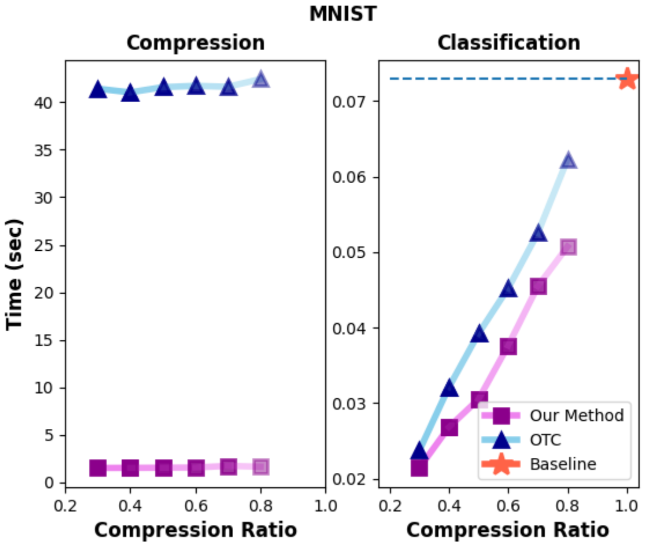

Figure 1: *MNIST Compression and Classification time comparison* In the Compression comparison figure on the left, the difference between OTC and our compression time is remarkable on MNIST. Also, for the smallest compression ratio with the best performance the classification time is less than half of baseline classification time.

2 FURTHER EMPIRICAL STUDY ON ρ

$$\rho := \mathbb{E}_{(X_V,C)\sim D}\left[\frac{P(X_V|C)}{P(X_V)}\right]$$

where $P(X_V)$ is the marginal probability and $P(X_V|C)$ is the conditional probability of node attributes given class variable i.e.

$$P(X_V|C) = \sum_c P(X_V|C=c)\pi(c) \;\; \text{where} \;\; \pi(c) \text{ is prior probability of class } c$$

To clarify what it means to have ρs with different percentages, we should consider their multiplication with the compression ratio as the ratio of the nodes that we tag as sensitive. But these sensitive nodes may or may not overlap with the ones that we compress using optimal transport (OT) in Algorithm 1. The final result of both ρ and OT compressed graph would have the proportion out of nodes equal to the compression ratio. To illustrate this look at Figure 2.

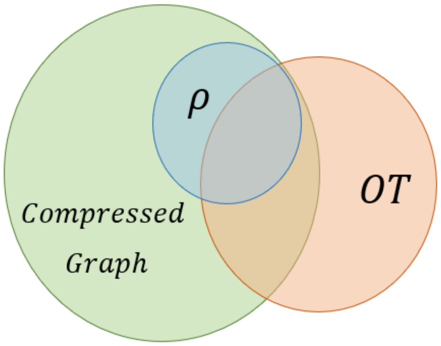

Figure 2: *Relation between compressed graph, sensitive nodes, and OT.* The final compressed graph should embrace all sensitive nodes, but might not have some OT without ρ.

Table 1: The table below the performance of our method with different ρ ratios for different compression ratios on CIFAR-10 dataset

ρ%	Acc@0.2	Acc@0.3	Acc@0.4	Acc@0.5	Acc@0.6	Acc@0.7	Acc@0.8
20	**0.507±0.031**	**0.522±0.033**	**0.536±0.032**	**0.53±0.034**	0.527±0.034	0.523±0.027	0.525±0.025
40	0.505±0.031	0.52±0.034	0.531±0.035	**0.53±0.034**	0.525±0.031	0.528±0.031	0.526±0.028
60	0.507±0.034	0.512±0.040	0.523±0.038	0.523±0.033	0.525±0.030	0.534±0.032	**0.528±0.026**
80	0.505±0.035	0.506±0.045	0.518±0.038	0.522±0.032	**0.528±0.030**	**0.536±0.030**	**0.528±0.026**

3 MI ESTIMATOR

An attributed graph consists of both a graph and a collection of node features. consider both the topology of the graph and a set of feature attributes that are attached to each node. In this section, we calculate the mutual information between an attributed graph and its class label, under certain conditional independence assumptions on the node attributes and edges. Let $g : (0, \infty) \to \mathbb{R}$ be a convex function with $g(1) = 0$. Given graph G_V with fixed set of vertices $V = \{v_1, \ldots, v_k\}$ and features set $X_V = \{X_{v_1}, \ldots, X_{v_k}\}$, the mutual information between (G_V, X_V) and class variable C with prior probability π_C is given by

$$I(G_V, X_V; C) = \mathbb{E}_{P_G, \pi_C}\left[g\left(\frac{P(G_V, X_V, C)}{P(G_V, X_V)\pi_C}\right)\right] = \mathbb{E}_{P_G, \pi_C}\left[g\left(\frac{P(G_V, X_V|C)}{P(G_V, X_V)}\right)\right], \quad (1)$$

where $P_G := P(G_V, X_V)$ and π_C is the prior probability of class C. We assume that the random node attributes are independent when conditioned on C – i.e. $P(X_V|C) = \prod\limits_{v \in V} P(X_v|C)$.

We also assume that conditioned on the class label and the attributes of the incident nodes, the indicator random variable for each edge is independent of everything else. In other words, conditioned on the class label, the graph is distributed according to a latent position vector model (see, e.g., Athreya et al. (2021)). In (1) $P(G_V, X_V|C) = P(X_V|C)P(G_V|X_V, C)$ and

$$P(G_V, X_V, C) = \pi_C P(G_V, X_V|C) = \pi_C P(X_V|C)P(G_V|X_V, C)$$

$$= \pi_C \cdot \prod_{v \in V} P(X_v|C) \cdot \prod_{\substack{E_{u,v} \in G_V \\ u,v \in V}} P(E_{u,v}|X_u, X_v, C) \cdot \prod_{\substack{E_{u,v} \notin G_V \\ u,v \in V}} (1 - P(E_{u,v}|X_u, X_v, C)). \quad (2)$$

In order to investigate information monotonicity, we first propose a new estimator for MI and apply it to summarized graphs.

Estimator $\widehat{I}(G_V, X_V; C)$:

Let $P_{u,v} := P(X_u, X_v)$ be the joint probability of random node features X_u and X_v for $u, v \in V$. Consider N i.i.d samples $\big\{(X_u, X_v)\big\}_{n=1}^{N}$ drawn from joint probability $P_{u,v}$ with adjacency values

$e_{u_1,v_1}, \ldots, e_{u_N,v_N}$ from adjacency matrices $A_1, \ldots, A_N$. Note that $e_{u,v}$ takes values either 0 or 1. We define a dependence graph $G^{(0)}$ as a directed multi-partite graph, consisting of two sets of nodes W_u and Z_v, with cardinalities denoted as $|W_u|$ and $|Z_v|$ respectively and with the set of all edges $E_{G^{(0)}}$. We map each point in the sets $X_u = \{X_{u1}, \ldots, X_{uN}\}$, and $X_v = \{X_{v1}, \ldots, X_{vN}\}$, to the nodes in new sets W_u and Z_v respectively, using the hash function H. Then let $H(x) = H_2(H_1(x))$, where the vector valued hash function $H_1 : \mathbb{R}^d \mapsto \mathbb{Z}^d$ is defined $H_1(x) = [h_1(x), \ldots, h_1(x_d)]$, for $x = [x_1, \ldots, x_d]$ and $h_1(x_i) = \lfloor \frac{x_i+b}{\epsilon} \rfloor$, for a fixed $\epsilon > 0$, and random variable $b \in [0, \epsilon]$. The random hash function $H_2 : \mathbb{Z}^d \mapsto \mathcal{F}$ is uniformly distributed on the output $\mathcal{F} = \{1, 2, \ldots, F\}$ where for a fixed tunable integer c_H, $F = c_H N$. Define

$$N^e_{i_u j_v} = \#\{(X_{u_n}, X_{v_n}) \;\; s.t. \;\; H(X_{u_n}) = i_u, H(X_{v_n}) = j_v \text{ and } e_{u_n,v_n} = 1\}, \tag{3}$$

which is the number of joint collisions of the nodes (X_{u_n}, X_{v_n}) at the pair (w_{i_u}, z_{j_v}). Let N_{i_u}, $N_{i_u j_v}$ be the number of collisions at the vertices (w_{i_u}), and (w_{i_u}, z_{j_v}) respectively, where

$$N_{i_u j_v} = \#\{(X_{u_n}, X_{v_n}) \;\; s.t. \;\; H(X_{u_n}) = i_u, H(X_v) = j_v\}, \tag{4}$$

By using $N^e_{i_u j_v}$, $N_{i_u j_v}$, and N_{i_u} we define $r^e_{i_u j_v} := \frac{N^e_{i_u j_v}}{N}$, $r_{i_u j_v} := \frac{N_{i_u j_v}}{N}$, $r_{i_u} := \frac{N_{i_u}}{N}$, and the following ratios,

$$r^e_{i_u j_v} := \frac{N^e_{i_u j_v}}{N}, \;\; r_{i_u j_v} := \frac{N_{i_u j_v}}{N}, \;\; r_{i_u} := \frac{N_{i_u}}{N}, \quad \text{and} \tag{5}$$

$$\widehat{P}_{i_u,j_v}(.|c) := \prod_{u\in V} r^c_{i_u} \cdot \prod_{u,v\in V} \frac{r^{e,c}_{i_u j_v}}{r^c_{i_u j_v}} \left(1 - \frac{r^{e,c}_{i_u j_v}}{r^c_{i_u j_v}}\right), \quad \widehat{P}_{i_u,j_v}(.) := \prod_{u\in V} r_{i_u} \cdot \prod_{u,v\in V} \frac{r^e_{i_u j_v}}{r_{i_u j_v}} \left(1 - \frac{r^e_{i_u j_v}}{r_{i_u j_v}}\right) \tag{6}$$

where $r^c_{i_u}$ is r_{i_u} with class label c. Further $r^{e,c}_{i_u j_v}$ and $r^c_{i_u j_v}$ are $r^e_{i_u j_v}$ and $r_{i_u j_v}$ with samples from class label c. Let $p_c = \frac{N_c}{N}$, where N_c is total number of sample with class label c. We propose a Hash-based estimator of $I(G_V, X_V; C)$ in (1) denoted by $\widehat{I}(G_V, X_V; C)$ as follows:

$$\widehat{I}(G_V, X_V; C) = \sum_c p_c \cdot \sum_{i_u, u\in V} \widehat{P}_{i_u,j_v}(.) \cdot g\left(\frac{\widehat{P}_{i_u,j_v}(.|c)}{\widehat{P}_{i_u,j_v}(.)}\right). \tag{7}$$

In particular, we use $g(x) = x \log x$. Note that $\sum\limits_{i_u, u\in V} := \sum\limits_{i_{v_1}, i_{v_2}, \ldots i_{v_{|V|}}} = \sum\limits_{i_{v_1}} \sum\limits_{i_{v_2}} \ldots \sum\limits_{i_{v_{|V|}}}$, where $|V|$ is cardinality of vertices of . Summed in (7) is over all edges in dependence graph $G^{(0)}$ having non-zero ratios. An important point on the proposed estimator is that for all vertex pairs $u, v \in V$ we have $r^{e,c}_{i_u j_v} \neq 0$ and $r^e_{i_u j_v} \neq 0$. This is coincident with using a dependence graph in MI estimator as the nodes and edges with zero collisions do not show up in the dependence graph. In practice if there exist a pair (u, v) such that $N^e_{i_u, j_v} = 0$, we eliminate the node from collision counts. Note that this estimator is inspired by Noshad et al. (2019) and the convergence rate for this estimator will be investigated in the future as it requires a fundamental study.

4 INFORMATION MONOTONICITY VIOLATION

The theoretical ramification of information non-monotonicity is that there exist certain data distributions for which the MI measure does not monotonically increase as the flow cost decreases. Empirically, it is important to investigate how common these distributions are (in other words, do they arise in practice?), how badly non-monotone they are, and how much this affects classification test accuracy. We leave the answers to these questions to the future but to initiate the investigation we ran an experiment to compare MI defined in (1) between Ours and OTC methods.

Figure 3 demonstrates the MI values using the MI estimator (described above in Section 3) and shows that our method outperforms OTC for smaller compression ratios but not for large ones which is a potential future investigation. We applied noise injection that is averaging over features of graph node neighbors randomly to provide a complex classification problem. We observe that our method outperforms OTC over accuracy and has higher MI after compression, however, this is slightly violated when the compression starts to grow.

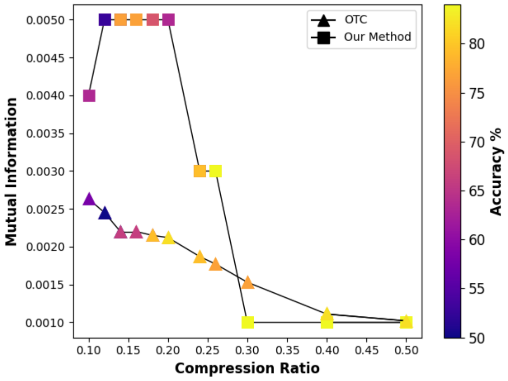

Figure 3: *MI Comparison for Synthetic Dataset* MI is shown for different compression ratios with each point color showing the performance on Synthetic data (100 samples and 50 nodes)

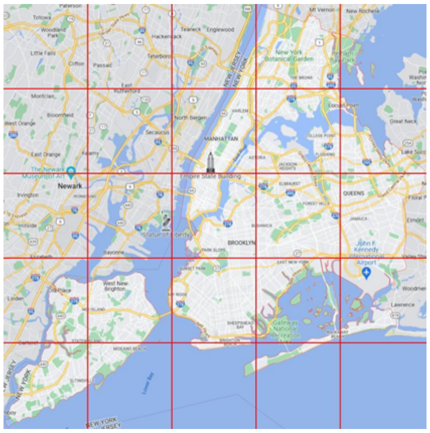

Figure 4: *NYC Grid map* Each square of the grid shows the nodes of graphs in NYC dataset. The attributes of the nodes are the total time of the trips inside that region. There is an undirected edge between two Nodes if there was any trip between them.

REFERENCES

Avanti Athreya, Minh Tang, Youngser Park, and Carey E. Priebe. On Estimation and Inference in Latent Structure Random Graphs. Statistical Science, 36(1):68 – 88, 2021. doi: 10.1214/20-STS787. URL `https://doi.org/10.1214/20-STS787`.

Morteza Noshad, Yu Zeng, and Alfred O Hero. Scalable mutual information estimation using dependence graphs. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2962–2966. IEEE, 2019.