Supplementary Material

Qi Zhao[*] **Ze Ye**[+] **Chao Chen**[+] **Yusu Wang**[*]
[*] Department of Computer Science and Engineering, The Ohio State University
[+] Department of Biomedical Informatics, Stony Brook University

1 Proof of Theorem 1

Theorem 1. *Let $G = (V, E)$ be a random graph sampled from* $\mathrm{SBM}(p_1, p_2, q, n_1, n_2)$*. Let $u \in C_1$ and $v \in C_2$, compute $\mathrm{Dg}G_u$ and $\mathrm{Dg}G_v$ as described above. Given any constant $\epsilon > 0$, the following two inequalities hold with high probability:*

$$\begin{aligned} W_1(\mathrm{Dg}_0 G_u, \mathrm{Dg}_1 G_v) \geq \\ c \cdot max\{n_1|p_1 - q - 2\epsilon|, n_2|p_2 - q - 2\epsilon|\} \\ W_1(\mathrm{Dg}_1 G_u, \mathrm{Dg}_1 G_v) \geq \\ c \cdot max\{n_1^2|p_1^3 - p_1 q^2 - 2\epsilon|, n_2^2|p_2^3 - p_2 q^2 - 2\epsilon|\} \end{aligned} \tag{1}$$

Proof. Denote the number of nodes in G_u from $C_i (i = 1, 2)$ as $N_i (i = 1, 2)$, the number of edges in G_u connecting two nodes from the same community C_i except u as N_{ii}, and the number of edges across two communities as N_e, then by Hoeffding's inequality, with any constant $\epsilon > 0$ we can conclude that

$$\begin{aligned} Pr(|N_1 - n_1 p_1| \leq \epsilon n_1) \geq& 1 - 2exp(-2\epsilon^2 n_1) \\ Pr(|N_2 - n_2 q| \leq \epsilon n_2) \geq& 1 - 2exp(-2\epsilon^2 n_2) \\ Pr(|N_{11} - n_1^2 p_1^3| \leq \epsilon n_1^2) \geq& 1 - 2exp(-2\epsilon^2 n_1^2)) \\ Pr(|N_{22} - n_2^2 q^2 p_2| \leq \epsilon n_2^2) \geq& 1 - 2exp(-2\epsilon^2 n_2^2) \\ Pr(|N_e - n_1 n_2 p_1 q^2| \leq \epsilon n_1 n_2) \geq& 1 - 2exp(-2\epsilon^2 n_1 n_2)) \end{aligned} \tag{2}$$

Consider 0-dim persistent homology, since all nodes in G_u are connected to u, the death of all topological features is 0. Recall the algorithm used for deriving our descriptor function (Eldridge et al., 2016), there are N_1 persistence points locating at the segment between $(p_1 - \delta, 0)$ and $(p_1 + \delta, 0)$ and N_2 points in the segment between $(p_2 - \delta, 0)$ and $(p_2 + \delta, 0)$ with probability μ.

Consider 1-dim extended persistent homology, we category cycles in G_u into 3 sets: $CS_1 = \{(u, u', u'')|u', u'' \in C_1\}$, $CS_2 = \{(u, u', u'')|u', u'' \in C_2\}$ and the set of all other cycles, CS_3. Each cycle in CS_1 or CS_2 leads to a persistence point in $\mathrm{Dg}_1 G_u$ lying in the segment between $(p_1 - \delta, 0)$ and $(p_1 + \delta, 0)$ or $(p_2 - \delta, 0)$ and $(p_2 + \delta, 0)$ with probability μ. Cycles in CS_3 corresponds to points along the diagonal with deviation of 2δ or within $[p_1 - \delta, p_1 + \delta] \times [q - \delta, q + \delta] \cup [p_2 - \delta, p_2 + \delta] \times [q - \delta, q + \delta]$ with the same probability. The size of CS_1 and CS_2 are determined by N_{11}, N_{22} and N_e. If $p_1 > p_2$, then $|CS_1| = N_{11} + N_e$ and $|CS_2| = N_{22}$. Otherwise $|CS_1| = N_{11}$ and $|CS_2| = N_{11} + N_e$.

The numbers of such nodes and edges are induced analogously within G_v. Specially, denote the number of nodes in G_v from $C_i (i = 1, 2)$ as $N_i' (i = 1, 2)$, the number of edges connecting two nodes in the same community C_i except v as N_{ii}', it follows that

$$\begin{aligned} Pr(|N_{11}' - n_1^2 q^2 p_1| \leq \epsilon n_1^2) \geq& 1 - 2exp(-2\epsilon^2 n_1^2)) \\ Pr(|N_{22}' - n_2^2 p_2^3| \leq \epsilon n_2^2) \geq& 1 - 2exp(-2\epsilon^2 n_2^2) \end{aligned} \tag{3}$$

The computation of $1 - th$ Wasserstein distance between $\mathrm{Dg}_0 G_u$ and $\mathrm{Dg}_0 G_v$ is somewhat counting persistence points in the two segments, diagonal and rectangle regions mentioned above. We focus on those lying along two segments. Clearly, the distance between two persistence points in the same segment is smaller than 2δ. After pairing persistence points from $\mathrm{Dg}_0 G_u$ and $\mathrm{Dg}_0 G_v$ in the same segment, the persistence points left unpaired and the diagonals are paired by a bijection function minimizing their summation distance. After given p_1, p_2 and q, the distance among points in different segments or diagonal are lower bounded by

$$c = min\{|p_1 - p_2|, |p_1 - q|, |p_2 - q|, q/\sqrt{2}\} \tag{4}$$

The number of persistence points unpaired with points lying in the same segment is $|N_1 - N_1'| + |N_2 - N_2'|$, it follows that

$$\begin{aligned} W_1(\mathrm{Dg}_0 G_u, \mathrm{Dg}_1 G_v) \geq \\ c \cdot max\{n_1|p_1 - q - 2\epsilon|, n_2|p_2 - q - 2\epsilon|\} \end{aligned} \tag{5}$$

with probability $(1 - 2e^{-2\epsilon^2 n_1})^2 (1 - 2e^{-2\epsilon^2 n_1})^2 \mu$.

The computation of 1-dim persistence diagrams Wasserstein distance follows the same method. Notice that persistence points locating within two rectangle

Table 1: Statistics of experimental benchmark datasets

	#Classes	#Features	#Nodes	#Edges	Edge density	Label rate
Cora	7	1433	2708	5429	0.0014	0.036
Citeseer	6	3703	3327	4732	0.0008	0.052
Pubmed	3	500	19717	44338	0.0002	0.003
Coauthor-CS	15	6805	18333	81894	0.0005	0.016
Coauthor-Physics	5	8415	34493	247962	0.0005	0.003
Amazon-Computers	10	767	13381	245779	0.0027	0.015
Amazon-Photo	8	745	7487	119043	0.0042	0.021

Table 2: Classification Accuracies on Benchmark Datasets

Method	**Cora**	**Citeseer**	**PubMed**	**Coauthor CS**	**Coauthor Physics**	**Amazon Computer**	**Amazon Photo**
PEGN-JI-2	82.5±0.5	71.7±0.6	78.7±0.6	92.7±0.3	94.1±0.3	84.0±1.0	92.2±0.5
PEGN-JI-1	82.4±0.5	71.7±0.5	78.5±0.6	92.7±0.3	94.1±0.2	86.1±0.6	92.7±0.4

regions can be paired as well. The number of persistence points unpaired with points lying in the same segment is $|N_{11} - N'_{11}| + |N_{22} - N'_{22}|$, which leads to

$$W_1(\mathrm{Dg}_1 G_u, \mathrm{Dg}_1 G_v) \geq c \cdot max\{n_1^2|p_1^3 - p_1 q^2 - 2\epsilon|, n_2^2|p_2^3 - p_2 q^2 - 2\epsilon|\} \tag{6}$$

with probability $(1 - 2e^{-2\epsilon^2 n_1^2})^2 (1 - 2e^{-2\epsilon^2 n_1^2})^2 \mu$.

□

2 Introduction of Datasets

Cora, Citeseer and Pubmed are citation graphs in which nodes represent documents and edges represent the undirected citation relations. Node features are elements of a bag-of-words representations of documents. In two Coauthor graphs, nodes represent authors which are connected by an edge if they jointly authored a paper. Node features are keywords for each author's papers, and node class labels are given by the authors' most active study fields. In two Amazon graphs, nodes represent goods and two nodes are connected if consumers frequently buy them together. Node features are bag-of-words encoded product reviews, and class labels indicate the product category. See Table 1 for the statistics of these datasets.

3 More Experiments Results

Besides Ollivier Ricci curvature, we also make experiments by taking Jaccard index as the weight function for graphs and construct subgraphs by picking 1-hop and 2-hop neighbourhoods around each node. Denote them **PEGN-JI-1** and **PEGN-JI-2**, respectively. The results are reported in Table 2.

References

Justin Eldridge, Mikhail Belkin, and Yusu Wang. Graphons, mergeons, and so on! In *Advances in Neural Information Processing Systems*, pages 2307–2315, 2016.