

A Supplementary Materials

A.1 Description of \mathcal{G} -Mixup hyperparameters

We describe the hyperparameters relevant to the \mathcal{G} -Mixup method. We also state the values for these hyperparameters used in the original paper, as stated in the paper; we note that some of these values in the original author’s code were different, and so we therefore changed it in our code.

- Augmentation ratio α : Multiplying the number of training graphs by this ratio gives the number of synthetic graphs generated by applying \mathcal{G} -Mixup to the graphons estimated from the training data. The authors use $\alpha = 0.2$ in their experiments.
- Mixup ratio λ and mixup interval $[\Lambda_1, \Lambda_2]$: Given graphons W_1, W_2 and mixup ratio $\lambda \in [0, 1]$, the authors’ algorithm produces a new mixed-up graphon

$$\lambda W_1 + (1 - \lambda) W_2. \quad (3)$$

The mixup ratio λ is randomly sampled from the interval $[\Lambda_1, \Lambda_2]$. The authors use the mixup interval $[0.1, 0.2]$ in their experiments.

- Augmentation number n_{aug} : In the \mathcal{G} -Mixup algorithm, we generate n_{aug} mixed-up graphons from any two distinct classes among all of the classes; the mixup ratios of these mixed-up graphons are given by λ_i for $i = 1, 2, \dots, n_{\text{aug}}$. We generate $\lfloor \alpha n / n_{\text{aug}} \rfloor$ synthetic graphs from each mixed-up graphon, where n is the size of the original training set. For binary classification with $\Lambda_1 = \Lambda_2$, this parameter is made irrelevant. The authors use $n_{\text{aug}} = 10$ in their experiments.
- Graphon resolution: If n is the graphon resolution, then in the graphon estimation step, an $n \times n$ matrix is used to represent the graphon. To sample from the graphon, an $n \times n$ adjacency matrix is then generated. Isolated nodes are then removed to provide a synthetic graph. Thus, the resolution influences the size of synthetic graphs, but does not determine it. In particular, it is an upper bound on the number of nodes in the synthetically generated graphs. The authors use the median number of nodes in the training set as the resolution.

A.2 Theoretical result details

We first provide some of the relevant definitions and background for Section 4.1.1 (Result 1).

Definition A.1. A **discriminative motif** F_G of a graph G is the subgraph with the minimal number of nodes and edges that can decide the class of the graph G . Let \mathcal{F}_G denote the set of discriminative motifs for a set \mathcal{G} of graphs.

The authors use this notion of discriminative motifs as their measure for “key topologies” among graphs of a certain class or label.

Given an arbitrary graph or graphon, we want to measure “how often” such a motif appears in the graph or graphon. The following definition will be used as this measure.

Definition A.2. Let F be an arbitrary graph.

- Let G be a graph. Then the **homomorphism density** of F with respect to the graph G is

$$t(F, G) = \frac{\text{hom}(F, G)}{|V(G)|^{|V(F)|}},$$

where $\text{hom}(F, G)$ denotes the total number of graph homomorphisms from F to G , and $|V(F)|, |V(G)|$ denote the number of nodes of the graphs F, G respectively.

- Let W be a graphon. Then the **homomorphism density** of F with respect to the graphon W is

$$t(F, W) = \int_{[0,1]^{|V(F)|}} \prod_{(i,j) \in E(F)} W(x_i, x_j) \prod_{i \in V(F)} dx_i,$$

where $E(F), V(F)$ denote the edge and vertex sets of F respectively.

Finally, the following norm on graphons provides a way to measure the “similarity” of graphons.

Definition A.3. Let $W : [0, 1]^2 \rightarrow \mathbb{R}$ be a measurable function. Then the **cut norm** of W is defined as

$$\|W\|_{\square} = \sup_{S, T \subseteq [0,1]} \left| \int_{S \times T} W(x, y) dx dy \right|,$$

where the supremum is taken over all measurable subsets $S, T \subseteq [0, 1]$.

We note that the use of the box (\square) in the cut norm notation is the standard notation, and is used to distinguish cut norm from other norms.

Remark A.4. The cut norm is indeed a norm. In particular, it is homogeneous, so for $\alpha \in \mathbb{R}$ is a scalar and $W : [0, 1]^2 \rightarrow \mathbb{R}$ a measurable function, then αW is also a measurable function on $[0, 1]^2$, and

$$\|\alpha W\|_{\square} = |\alpha| \|W\|_{\square}.$$

In particular, note that

$$\|W\|_{\square} = \|-W\|_{\square}.$$

We now provide our detailed proof of Lemma 4.1.

Proof of Lemma 4.1. We expand upon the proof of this result in [2, Lemma 4.1].

First, notice that an equivalent definition for the cut norm of a graphon U is

$$\|U\|_{\square} = \sup_{f, g} \left| \int_{[0,1]^2} U(x, y) f(x) g(y) dx dy \right|, \quad (4)$$

where the supremum is taken over all measurable functions $f, g : [0, 1] \rightarrow [0, 1]$.

Let $V(F) = \{1, \dots, n\}$ and $E(F) = \{e_1, \dots, e_m\}$, where $e_t = (i_t, j_t)$. Then

$$t(F, W) - t(F, W') = \int_{[0,1]^n} \left(\prod_{e_t \in E(F)} W(x_{i_t}, x_{j_t}) - \prod_{e_t \in E(F)} W'(x_{i_t}, x_{j_t}) \right) \prod_{i \in V(F)} dx_i.$$

For $t = 1, \dots, m$, define

$$\begin{aligned} X_t(x_1, \dots, x_n) &= \left(\prod_{k=1}^{t-1} W(x_{i_k}, x_{j_k}) \right) (W(x_{i_t}, x_{j_t}) - W'(x_{i_t}, x_{j_t})) \left(\prod_{k=t+1}^m W'(x_{i_k}, x_{j_k}) \right) \\ &= W(x_{i_1}, x_{j_1}) \cdots W(x_{i_t}, x_{j_t}) W'(x_{i_{t+1}}, x_{j_{t+1}}) \cdots W'(x_{i_m}, x_{j_m}) \\ &\quad - W(x_{i_1}, x_{j_1}) \cdots W(x_{i_{t-1}}, x_{j_{t-1}}) W'(x_{i_t}, x_{j_t}) \cdots W'(x_{i_m}, x_{j_m}). \end{aligned}$$

Notice that for $t = 1, \dots, m-1$, the second term of $X_t(x_1, \dots, x_n)$ cancels out with the first term of $X_{t+1}(x_1, \dots, x_n)$. Thus,

$$\prod_{e_t \in E(F)} W(x_{i_t}, x_{j_t}) - \prod_{e_t \in E(F)} W'(x_{i_t}, x_{j_t}) = \sum_{t=1}^m X_t(x_1, \dots, x_n).$$

Fix all variables $x_j \in [0, 1]$, where $k \neq i_t, j_t$; we then have that

$$f(x_{i_t}) = \prod_{k=1}^{t-1} W(x_{i_k}, x_{j_k}), \quad g(x_{j_t}) = \prod_{k=t+1}^m W'(x_{i_k}, x_{j_k})$$

are both measurable functions $[0, 1] \rightarrow [0, 1]$ in x_{i_t} and x_{j_t} respectively. Then by Equation 4,

$$\left| \int_{[0,1]^2} X_t(x_1, \dots, x_n) dx_{i_t} dx_{j_t} \right| \leq \|W - W'\|_{\square},$$

and so

$$\begin{aligned} \left| \int_{[0,1]^n} X_t(x_1, \dots, x_n) \prod_{i=1}^n dx_i \right| &\leq \int_{[0,1]^{n-2}} \left| \int_{[0,1]^2} X_t(x_1, \dots, x_n) dx_{i_t} dx_{j_t} \right| \prod_{i \neq i_t, j_t} dx_i \\ &\leq \|W - W'\|_{\square}. \end{aligned}$$

Thus, we see that

$$\begin{aligned} |t(F, W) - t(F, W')| &= \left| \int_{[0,1]^n} \sum_{t=1}^m X_t(x_1, \dots, x_n) \prod_{i=1}^n dx_i \right| \\ &\leq \sum_{t=1}^m \left| \int_{[0,1]^n} X_t(x_1, \dots, x_n) \prod_{i=1}^n dx_i \right| \\ &\leq m \|W - W'\|_{\square}. \end{aligned}$$

Recall that $m = e(F)$, and so we have the desired inequality. \square

A.3 Figures for experimental results

The following figures show the loss curves for experiments from Sections 4.1.2 and 4.2.1, as well as the original paper's loss curves.

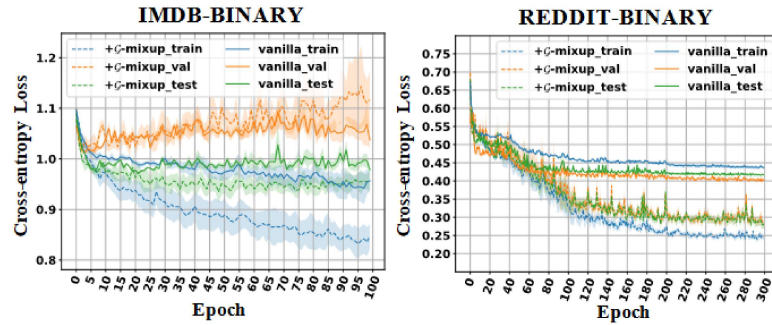


Figure 1. The training/validation/test loss curves on IMDB-B and REDDIT-B with GCN as backbone. The curves are depicted on ten runs. This is part of Figure 4 in the original paper [1].

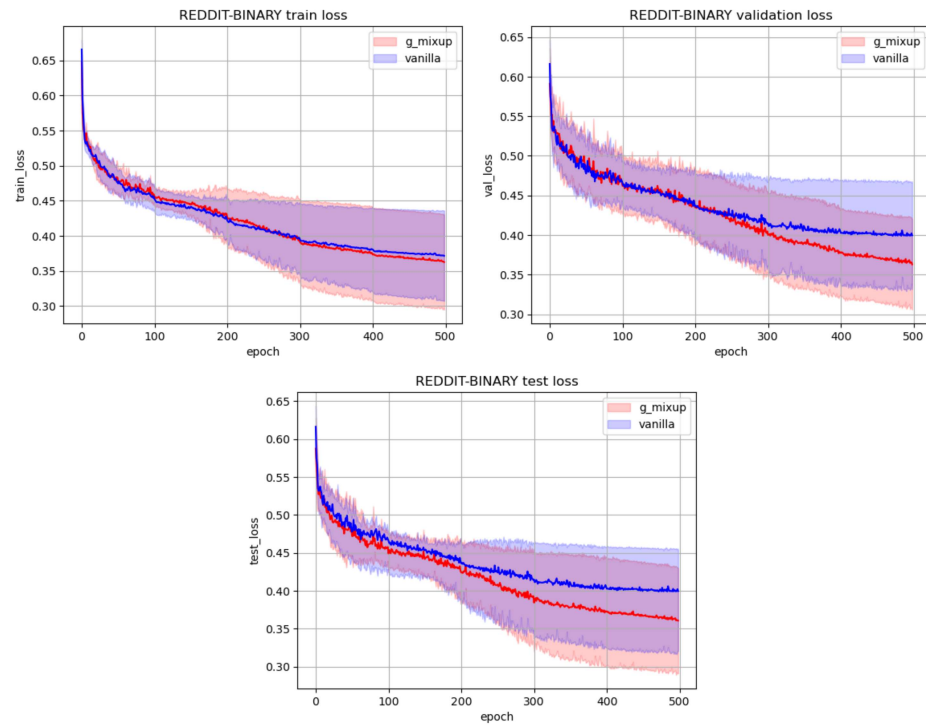


Figure 2. The training/validation/test curves on REDDIT-B with GCN as a backbone. The curves are depicted on ten runs. The line is the mean of the corresponding loss, while the shaded area is \pm the standard deviation of the losses.

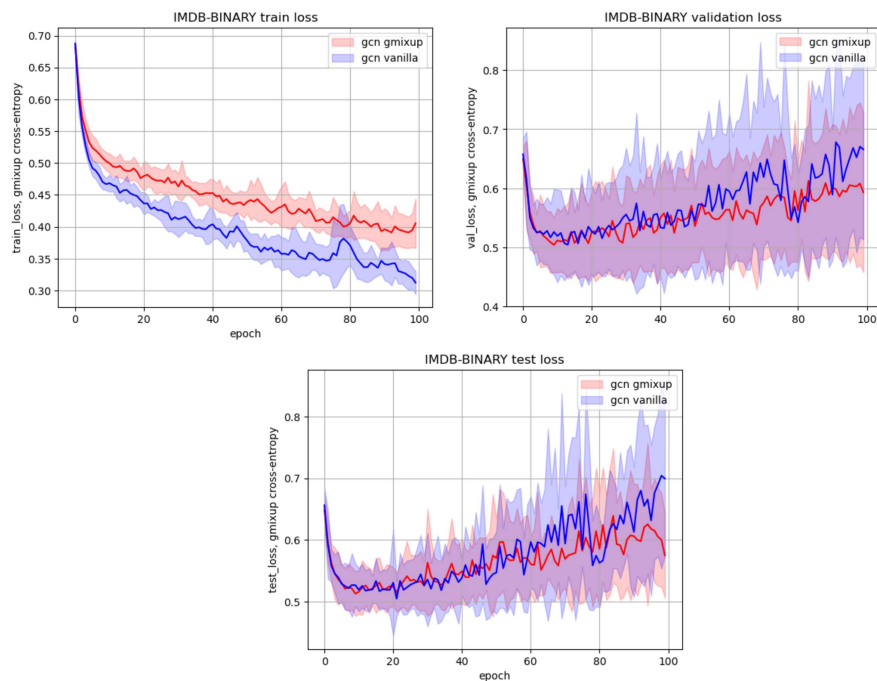


Figure 3. The training/validation/test curves on IMDB-BINARY with GCN as a backbone. The curves are depicted on ten runs. The line is the mean of the corresponding loss, while the shaded area is \pm the standard deviation of the losses.

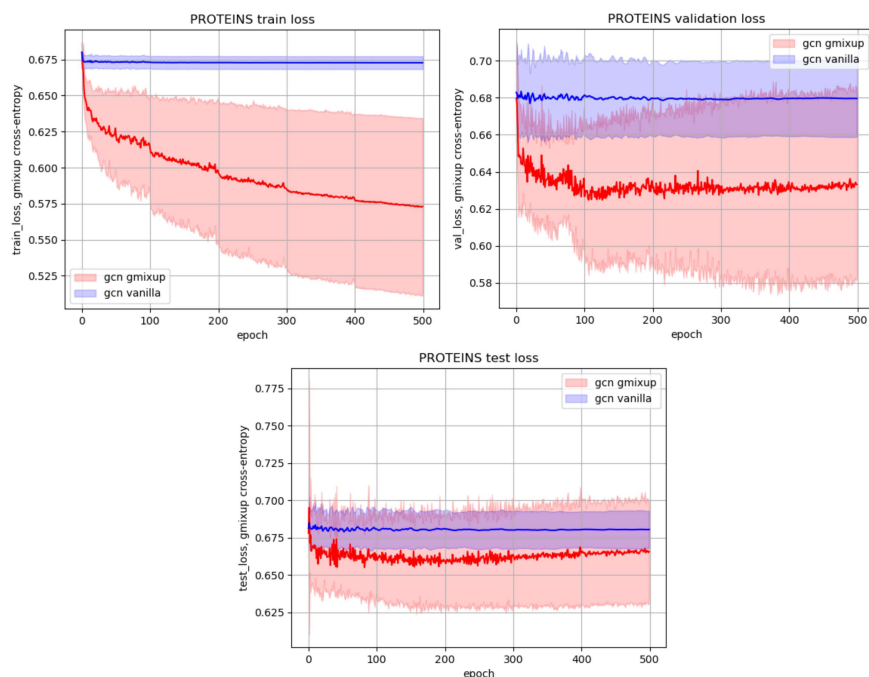


Figure 4. The training/validation/test curves on PROTEINS with GCN as a backbone. The curves are depicted on ten runs. The line is the mean of the corresponding loss, while the shaded area is \pm the standard deviation of the losses.