Graph Representation Learning

Please use the official LaTeX template to type your answers available in Moodle. Please respect the notation from Table 1 in your answers whenever applicable:

Table 1: Notation.

σ	An element-wise non-linearity.
t	Iteration, or layer t.
$d^{(t)}$	The dimension of a vector at iteration t.
d	The dimension of a vector, and an abbreviation for $d^{(0)}$.
$\mathbf{1}^d \in \mathbb{R}^d$	A d-dimensional vector of all 1's.
$\mathbb{I}^d \subseteq \mathbb{R}^d$	The set of d-dimensional one-hot vectors.
$\mathbb{B}$	Boolean domain $\{0, 1\}$.
$\mathbf{b}^{(t)} \in \mathbb{R}^d$	A bias vector.
$\mathbf{x}_u \in \mathbb{R}^d$	The feature of a node $u \in V$.
$\mathbf{h}_u^{(t)} \in \mathbb{R}^{d^{(t)}}$	The representation of a node $u \in V$ at layer t.
$\mathbf{z}_u = \mathbf{h}_u^{(T)} \in \mathbb{R}^{d^{(T)}}$	The final representation of a node $u \in V$ after T layers/iterations.
$\mathbf{W}_x^{(t)} \in \mathbb{R}^{d^{(t+1)} \times d^{(t)}}$	Learnable parameter matrix at layer t.
MLP	A multilayer perceptron with ReLU as nonlinearity.

Question 3

In order to train expressive Graph Neural Networks (GNNs), model architectures need to be able to distinguish graphs even at initialisation time. In their recent work on *Zero-One Laws of Graph Neural Networks*[1], Adam-Day et al. formally and experimentally proved the negative result that, under mild assumptions, binary classification models, such as Graph Convolutional Networks (GCNs), obey a 0-1 Law on Erdös-Rényi graphs. The 0-1 Law asserts that, as the graphs size increases, the model output collapses to a fixed binary prediction, regardless of its input graph. In practice, the authors' approach relies on proving that the model converges to a fixed average graph embedding vector, which is then a sufficient condition for the 0-1 Law to hold true. We name this condition the convergence hypothesis $\mathcal{H}$, and, in this report, we ask the following question: *In concrete experimental settings, does the 0-1 Law only hold true when the convergence hypothesis $\mathcal{H}$ is satisfied, or does it generalize to additional settings?*
The objectives of this study include:

1. Highlighting the convergence hypothesis $\mathcal{H}$ for the GCN and Mean GNN models, since $\mathcal{H}$ was theoretically proven for the general theorem but not empirically shown in Adam-Day et al[1].
2. Examining the convergence hypothesis $\mathcal{H}$ on data that was not generated from the Erdös-Rényi random graphs model.
3. Examining $\mathcal{H}$ on a set of models beyond those explored in Adam-Day et al[1].

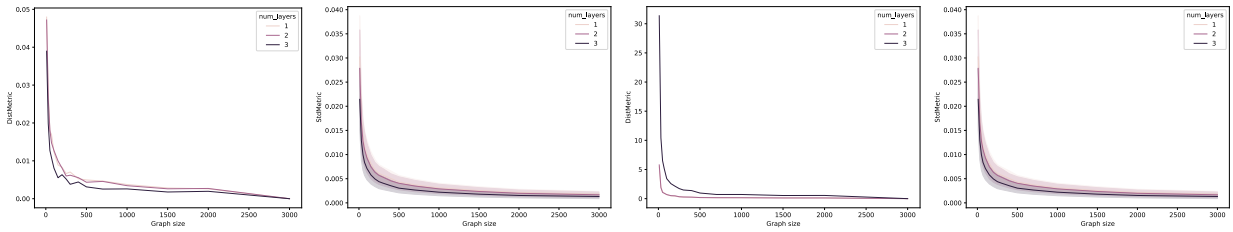
The code for the experiments is available online at github.com/grl-student/grl-mini-project. Our implementation closely aligns with the choices outlined in Adam-Day et al.[1]. Random graphs are generated once and for all for sizes ranging from 10 to 3000 nodes, with 30 graphs per graph size. Experiments encompass the following hyperparameters, ensuring that the necessary conditions for the 0-1 Law theorem are met:

- The core architecture is followed by an average pooling and by a 2-layer MLP, with a ReLU activation function. The final activation function is the sigmoid function. Layers are initialized according to the Xavier-Glorot initialization. We also use the ReLU activation function for the graph architecture, as it is Lipschitz continuous.
- 10 models are evaluated on 30 generated graphs for each graph dimension. The uniform (sub-Gaussian) distribution is used to generate node features of dimension 128. The edge rate is set to 0.5. Models are randomly initialised and not trained.

To experimentally prove hypothesis $\mathcal{H}$, we use two metrics to measure the convergence of the average embedding vector. The first one, termed DistMetric, computes the L2-norm between the average graph embedding at a fixed graph dimension and the limit vector for an infinitely large graph. Since the limit vector is not known, we approximate it by its average value at the highest graph dimension, namely 3000. For the second metric, we measure whether the vector is converging by computing the standard deviation across all graphs of a given size, and checking whether this statistic converges to zero across all coordinates as the size of the graph grows to infinity. This is referred to as the StdMetric in the rest of this report.

Experiment 1: GCN and Mean GNN models.

[1] *Zero-One Laws of Graph Neural Networks*, Adam-Day, Iliant and Ceylan. NeurIPS 2023

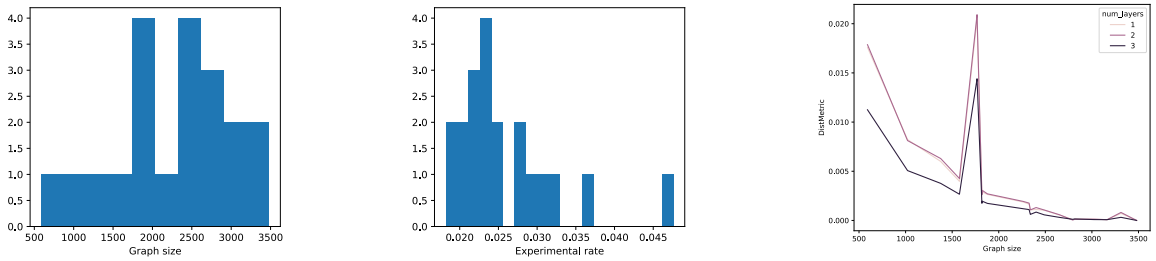


(a) (b) (c) (d)

Figure 5: DistMetric for GCN models and MeanGNN models in figures 5a and 5c resp., and StdMetric for GCN and MeanGNN models in figures 5b and 5d resp., as functions of graph size.

For both models, the embedding vector tends to a constant vector as graphs become larger. The result is independent of the number of layers in the model and remains true even for 1 layer (although the convergence rate depends on the number of layers). This result was implicitly covered by the proof of the convergence hypothesis $\mathcal{H}$ in the paper[1], but we confirm here that we can observe it empirically, even for a graph size smaller than 3000. This motivates the use of sizes up to 3000 for the rest of this report.

Experiment 2: GCN model on the PPI dataset. Random graphs are often not representative of real-world data, as they often exhibit specific structural properties that Erdös-Rényi random graphs may not capture (such as cycles or clusters). Having that in mind, we selected a real-world dataset whose graph sizes are of the order of that chosen for the previous experiments: the Protein-Protein Interaction (PPI) dataset[3]. Experiments are similar to the previous ones, using 10 GCN models with average pooling and from 1 to 3 layers.



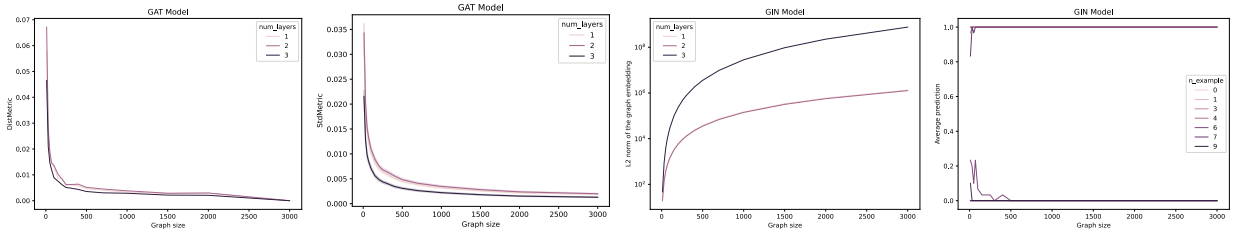
(a) (b) (c)

Figure 6: Histogram of graph sizes in the PPI dataset in figure 6a. Histogram of the experimental rate in the PPI dataset in figure 6b. The experimental rate is the number of edges divided by the maximum number of edges ($\frac{n(n-1)}{2}$). DistMetric as a function of the graph size in figure 6c.

Apart from the anomaly at around 1700 nodes, the graph embedding converges to a constant vector. All models exhibit a 0-1 Law, showing that this behaviour can be extended to other graph distributions. Figure 6b shows that the rate is not constant across all graphs in the dataset. This proves that the assumption of having graphs generated with Erdös-Rényi model is not necessary for the convergence hypothesis $\mathcal{H}$ to hold true. Therefore this collapsing behaviour may also occur in real-world datasets, although a more thorough investigation would be required for any conclusive statement on the empirical effect of the data distribution.

[3]https://paperswithcode.com/dataset/ppi

Experiment 3: GAT and GIN models. We now attempt to extend the 0-1 Law and to evaluate the convergence hypothesis $\mathcal{H}$ on the GAT and GIN models with $1, 2$ and 3 layers. Parameters are similar as previously, using 10 models each time.



(a) (b) (c) (d)

Figure 7: DistMetric and StdMetric for GAT models as functions of graph size, in figures 7a and 7b respectively. Norm of the graph embedding vector for GIN models as a function of the graph size, in figure 7c. Average prediction per graph size in figure 7d for models with 3 layers.

The GAT model follows the convergence hypothesis $\mathcal{H}$ as shown in figures 7*a* and 7*b*. However, in the case of the GIN model, the norm of the graph embedding vector grows exponentially as the graph size increases, showing that the average embedding vector does not converge to finite values, at least for the range of graph sizes studied here. This observation potentially invalidates the convergence hypothesis $\mathcal{H}$ in this setting. Nevertheless, the GIN model still appears to follow the 0-1 Law as shown in figure 7d. This is a potential example of a setting in which the convergence hypothesis $\mathcal{H}$ is not necessary for the 0-1 Law to hold true. The proof scheme developed in Adam-Day et al[1] would not be applicable here, calling for the exploration of alternative theoretical proofs that could extend and generalize the theorem introduced in their paper.

Discussion While the work of Adam-Day et al[1] focuses on the 0-1 Law, this report investigates the convergence hypothesis $\mathcal{H}$, which is a sufficient condition for the 0-1 Law to hold true. In particular, our preliminary experiments show that (1) the convergence hypothesis can hold true outside of the data distribution generated by the Erdös-Rényi random graphs model, and (2) the hypothesis empirically holds true for the GCN, Mean GNN and GAT models but not for the GIN model (at least in the regime considered here).

More interestingly, we observe that the 0-1 Law still holds approximately true for the GIN model, despite the lack of convergence of the average embedding vector. As a result, we hypothesize that the 0-1 Law may hold more generally than shown in Adam-Day, e.g. due to saturating effects of the activations rather than the convergence of the average embedding vector.

Conclusion These results show that randomly-initialized models are already limited in their expressiveness, and unable to distinguish graphs with a high number of nodes even before the training starts. It seems that the 0-1 Law is quite common, for various models and datasets, and many conditions stated in the original theorem could be relaxed. Experiment 2 showed similar behaviour could be observed with real-world data, impacting current attempts to develop GNNs.

In order to strengthen our experimental findings, we would need to use a greater variety of datasets, as well as exploring larger graph sizes to truly approximate the asymptotical behaviour. Further, in order to obtain statistically significant results, we would need to run a larger number of repeated experiments.