A Appendix

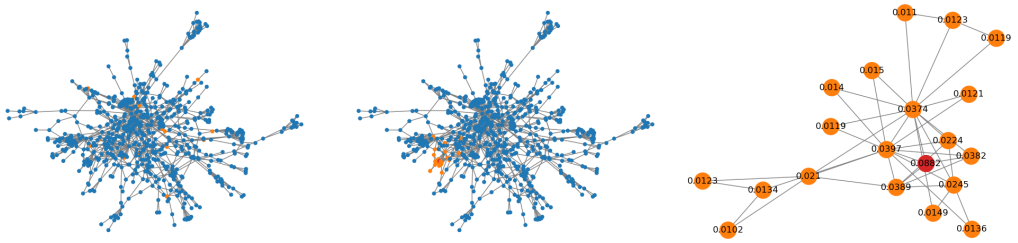
A.1 Data Statistics

Nodes in the graph are anonymized and do not contain any personally identifiable data.

Table 5: Overall Dataset Statistics

Dataset	# Nodes	# Edges	# Features	# Classes
Cora	2,708	5,278	1,433	7
Citeseer	3,327	4,614	3,703	6
PubMed	19,717	44,325	500	3
ogb-arxiv	169,343	1,166,243	128	40
Reddit	232,965	114,615,892	602	41

A.2 Scalable biased sampler details



(a) IID sample **(b)** Biased sample **(c)** PPR-score on biased sample

Figure 5: A biased sample on Cora dataset for one class, orange indicates the training data, red indicates the initial seed used in our PPR-S sampler. The PPR-score is presented in figure (c).

In the above example of our scalable personalized pagerank sampler (PPR-S), we set $\epsilon = 0.005$ and compute the top-γ approximated personalized page rank vector $\pi_i^{\text{ppr}}(\epsilon)$ for a randomly selected node i. If a seed node i has enough PPR-neighbors (non-zero entries), we add the top-γ PPR neighbors into the training set (γ=20). In Figures 5a and 5b, we visualize a sub-network of a specific class in Cora [27]; training samples are colored orange. Specifically, we visualize the nodes in biased sample and their PPR-score *w.r.t.* the seed node in color red in Figure 5c. The algorithm for biased training set creation is described in Algorithm 1.

Algorithm 1: Biased Training Set Creation PPR-S(γ, ϵ, α)

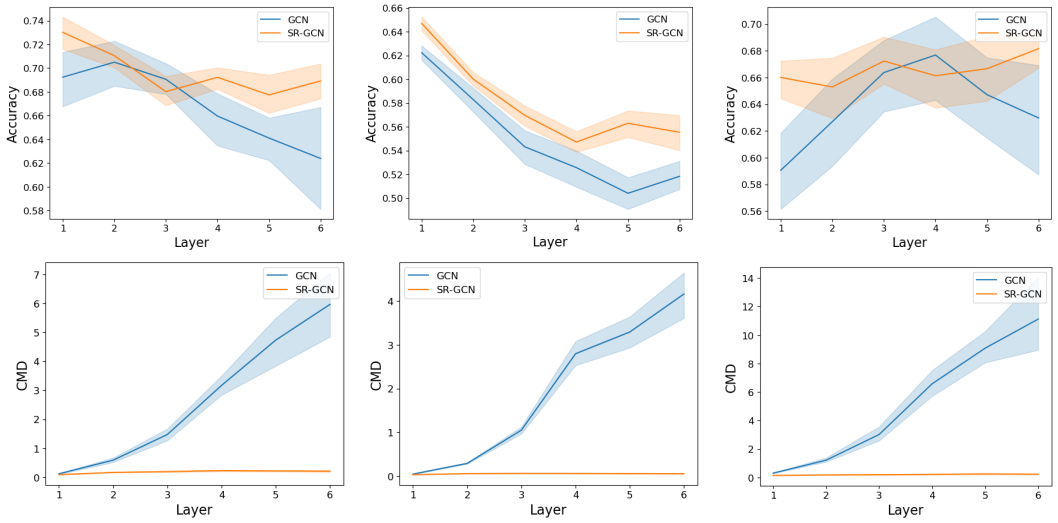
Given a class c, label ratio τ, graph size N;
Initialize the biased training set X = { } ;
while *len(X)*$< N \cdot \tau$ **do**
  Sample node i of class c, compute its top-γ entries in $\pi_i^{\text{ppr}}(\epsilon)$ via [2];
  if $\pi_i^{ppr}(\epsilon)$ *has* γ *non-zero entries* **then**
    X.add($\pi_i^{\text{ppr}}(\epsilon)$) ;
  end
end

A.3 SR-GNN with increasing model complexity

A.3.1 Performance of SR-GNN in deeper models

In Section 3.1 of the main paper, we highlighted that the graph inductive bias can amplify the 'normal' shift of non-IID samples. Therefore, in a deeper GNN, the negative influence of biased training data is expected to be larger. We present the accuracy (Micro-F1) and distribution shifts (CMD) of a deeper GCN [15] model and our Shift-Robust GCN using regularization proposed in Section 4.1 of the main paper. In Figure 6, the distribution shift dramatically increases as the depth of the model grows. On Cora and Citeseer, SR-GCN can effectively improve the performance regardless of the depth of the model. On PubMed, we observe that the performance of GCN increases when there are

fewer than 4 hidden layers. In most cases, SR-GCN still outperforms its base model with a smaller variance.

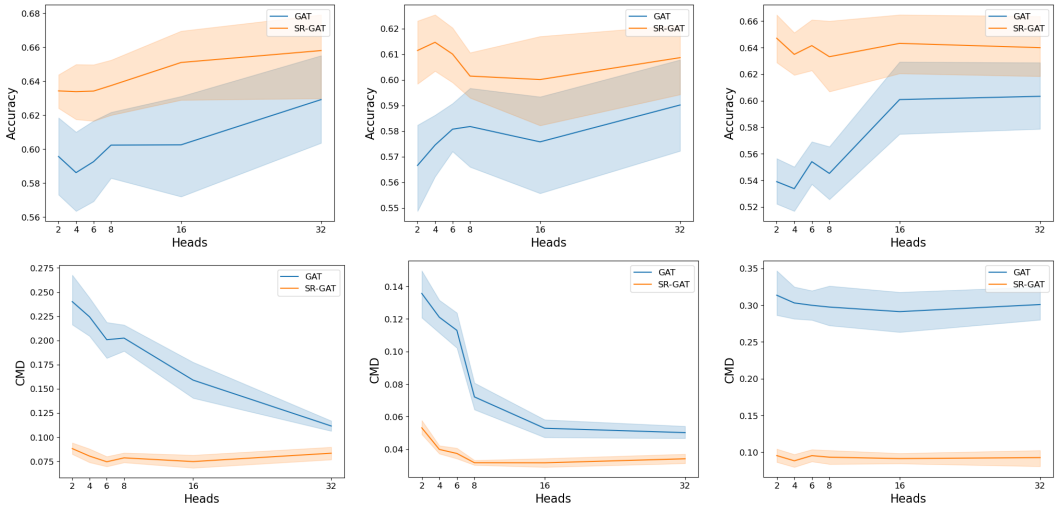


(a) Cora **(b)** Citeseer **(c)** Pubmed

Figure 6: Comparison of GCN vs. SR-GCN model performance with the the same parameters. Our shift-robust algorithm boosts the performance (top) consistently by reducing the distribution shifts (bottom).

A.3.2 Performance of SR-GNN in wider models

Besides the depth of the model, increasing the width of the model is another way to increase the complexity and capacity of the model. We vary the number of heads in GAT [31] while keeping the number of the GAT layers (*i.e.* 2) and hidden dimension of each attention head fixed (*i.e.* 32) and report back the performance of GAT and SR-GAT in Figure 7. In general, more attention heads lead to better performance and smaller distribution shifts (see lower figures). SR-GNN provides robust improvements across various number of the attention heads.



(a) Cora **(b)** Citeseer **(c)** Pubmed

Figure 7: Comparison of GAT vs. SR-GAT model performance under increasing attention heads. Our shift-robust algorithm boosts the performance (upper) consistently by reducing the distribution shifts (lower).

A.4 Performance study with best hyper parameters

In Section 5.4 of the main paper, we presented the parameter study of SR-GNN regarding the distribution discrepancy regularizer and instance weighting. We tune these parameters to obtain the best performance of SR-GNN **w.o.** IR, SR-GNN **w.o.** Reg. and SR-GNN in Table 6. Both SR-GNN and its variants show improvement performance and minimize the performance gap with the IID trained model even further.

Table 6: Comparison with ablations with best tuned parameters on three dataset.

Method	**Cora**			**Citeseer**			**PubMed**		
	Micro-F1↑	Macro-F1↑	ΔF1 ↓	Micro-F1↑	Macro-F1↑	ΔF1 ↓	Micro-F1↑	Macro-F1↑	ΔF1 ↓
GCN (IID)	80.8 ± 1.6	80.1 ± 1.3	0	70.3 ± 1.9	66.8 ± 1.3	0	79.8 ± 1.4	78.8 ± 1.4	0
w.o. IR.*	72.4 ± 3.5	70.1 ± 3.8	8.4	64.9 ± 1.3	62.4 ± 1.0	5.4	70.4± 3.5	68.8 ± 4.0	9.4
w.o. Reg.*	73.4 ± 2.7	71.1 ± 3.4	7.4	66.8 ± 1.1	64.0 ± 1.0	3.5	66.4 ± 4.0	64.0 ± 5.5	13.4
SR-GNN *	**74.3** ± **2.5**	**71.9**± **3.1**	**6.5**	**67.5** ± **1.3**	**64.4** ± **1.2**	**2.8**	**71.3** ± **2.2**	**70.2** ± **2.4**	**8.5**

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See conclusions.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Code included in supplemental material
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] These are widely used public datasets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] In appendix
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]