A DETAILS FOR REPRODUCIBILITY

Table 5: The details of hyper-parameters used in FT.

Model	Dataset	Temperature	λ
GCN	Cora	0.1	50
	Citeseer	4	0.1
	Pubmed	0.01	1
	A-Computers	12	50
	A-photo	0.001	200
GAT	Cora	0.01	50
	Citeseer	0.01	1
	Pubmed	4	200
	A-Computers	24	200
	A-Photo	12	200
GraphSAGE	Cora	4	1
	Citeseer	4	0.1
	Pubmed	1	0.01
	A-Computers	24	100
	A-Photo	12	100
APPNP	Cora	8	0.1
	Citeseer	0.1	0.5
	Pubmed	20	1
	A-Computers	24	100
	A-Photo	12	100
SGC	Cora	4	0.1
	Citeseer	4	0.5
	Pubmed	0.1	0.1
	A-Computers	12	50
	A-Photo	4	50

A.1 Experimental Environments

We run all our experiments on a single GPU device of GeForce GTX 1080 with 11 GB memory, and the operating system is Ubuntu 16.04.6. Besides, we implement our framework based on Deep Graph Library (DGL) of version 0.6.0 and Pytorch of version 1.8.1.

A.2 Brief Comments on Data Preparation

We use the largest connected component as previous works did [21, 31] and follow the dataset settings in [21, 31], except that we use 40/10 nodes instead of 20/30 per class for training/validation, because we think it is more reasonable to have more nodes in the training set. Hence the results of baselines in our experiments are different from those in their original papers. A brief introduction of the datasets is as follows:

• Citeseer and Cora [20] are two citation datasets consisting of scientific papers, where node features indicate whether words within the dictionary appear in the paper or not, and labels indicate the fine-grained research domain of each paper.

• Pubmed [18] is a citation dataset consisting of papers related to diabetes in the PubMed database, where node features are TF-IDF weighted word frequencies, and labels indicate the type of diabetes discussed in each article.

Table 6: The details of hyper-parameters used in LTD.

Teacher	Dataset	α	β	k	λ
GCN	Cora	1.92E-04	2.60E-04	1	1
	Citeseer	6.90E-06	7.46E-05	1	1
	Pubmed	7.40E-06	9.38E-05	1	200
	A-computers	3.85E-06	1.18E-05	1	100
	A-photo	1.01E-05	1.89E-04	1	50
GAT	Cora	1.68E-03	3.29E-04	2	1
	Citeseer	6.24E-05	1.71E-04	2	1
	Pubmed	2.95E-05	3.72E-06	2	200
	A-computers	3.56E-06	6.82E-05	2	200
	A-photo	1.89E-05	1.78E-04	2	50
GraphSAGE	Cora	8.74E-04	2.24E-04	3	1
	Citeseer	9.79E-06	2.06E-04	3	1
	Pubmed	5.58E-05	4.79E-04	3	100
	A-computers	6.86E-07	1.12E-05	3	200
	A-photo	7.19E-06	6.60E-05	3	50
APPNP	Cora	7.25E-04	4.15E-04	3	1
	Citeseer	2.60E-04	4.42E-05	3	1
	Pubmed	4.68E-05	7.51E-05	1	200
	A-computers	8.60E-06	1.42E-05	3	200
	A-photo	2.99E-05	5.64E-06	3	50
SGC	Cora	7.14E-04	1.92E-04	2	1
	Citeseer	1.24E-05	1.35E-04	1	1
	Pubmed	7.92E-05	4.16E-04	1	100
	A-computers	6.00E-06	8.52E-05	1	200
	A-photo	1.67E-05	8.40E-05	1	50

• A-Computers and A-Photo [21] are two co-purchase datasets consisting of products, where node features represent bag-of-words encoded product reviews, and labels indicate the product category.

A.3 Details of Teacher/Student Models

In our experiments, we use the following five representative GNN models as teacher/student models. All the pretrained GNN teachers are carefully tuned to guarantee that their performance cannot be improved via hyper-parameter tuning, and the the details are as follows:

• GCN [13]: we use a 1-layer GCN with hidden dimension as 64, drop probability as 0.8 and weight decay of Adam as 0.001.

• GAT [25]: we employ a 2-layer GAT with hidden dimension as 64, drop probability as 0.6, weight decay of Adam as 0.01, attention dropout probability as 0.3. Specifically, we use 8-head attention to enhance the model.

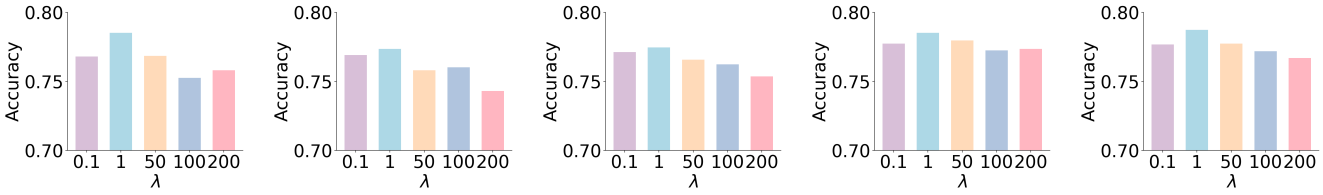
• GraphSAGE [9]: we set GraphSAGE with hidden dimension as 128, sample number as 5, batch size as 256, and weight decay of Adam as 0.0005. Specifically, we use GCN-based variant function as its aggregation function.

• APPNP [14]: we use a 2-layer APPNP with hidden dimension as 64, drop probability as 0.5, and weight decay of Adam as 0.01. Specifically, we use 10 power iteration steps.

• SGC [26]: we employ SGC with weight decay of Adam as 0.001 and use a 2-layer setting.

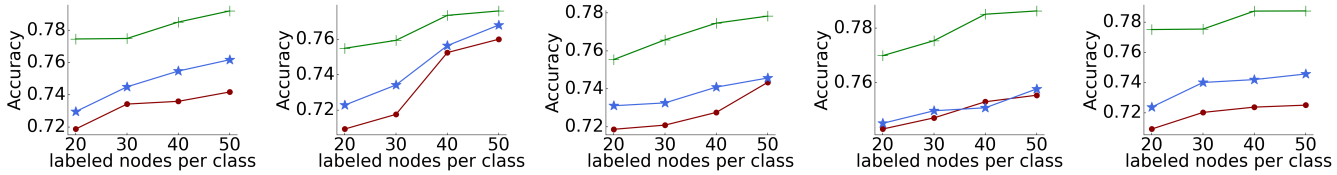
Table 7: Classification accuracies on training/validation/test set of A-Computers. The gap indicates the difference between the performance on validation and test sets.

Model	GNN Teacher				FT				LTD			
	Training	Validation	Test	Gap↓	Training	Validation	Test	Gap↓	Training	Validation	Test	Gap↓
GCN	0.920	0.930	0.859	0.071	0.948	0.940	0.847	0.093	0.915	0.920	0.865	**0.055**
GAT	0.795	0.890	0.809	0.081	0.848	0.900	0.808	0.092	0.815	0.880	0.830	**0.050**
SAGE	0.895	0.910	0.783	0.127	0.908	0.890	0.800	0.090	0.873	0.870	0.814	**0.056**
APPNP	0.910	0.900	0.810	0.090	0.965	0.930	0.814	0.116	0.888	0.910	0.836	**0.074**
SGC	0.950	0.890	0.805	0.085	0.983	0.920	0.833	0.087	0.730	0.850	0.853	**-0.003**



(a) GCN (b) GAT (c) GraphSAGE (d) APPNP (e) SGC

Figure 7: Classification accuracies under different balance hyper-parameter λs on Citeseer dataset.



(a) GCN (b) GAT (c) GraphSAGE (d) APPNP (e) SGC

Figure 8: Classification accuracies under different training ratios on Citeseer dataset. Green lines are LTD; Blue lines are FT; Red lines are the teacher models.

A.4 Settings for Other Distillation Frameworks

In our experiments, we use the following two knowledge distillation frameworks as baselines.

(1) CPF [31]: We train CPF in the inductive setting, with the number of mlp layers as 1. And we employ Optuna to explore the number of propagation layers K from $\{6, 7, 8, 9, 10\}$, global temperature from $\{0.001, 0.01, 0.1, 1, 4, 8, 12, 16, 20, 24\}$, hidden size in MLP from $\{8, 16, 32, 64\}$, dropout rate from $\{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, learning rate from $\{0.001, 0.005, 0.01\}$, and weight decay of Adam optimizer from $\{0.0005, 0.001, 0.01\}$.

(2) RDD [35]: We fix RDD with learning rate as 0.1 and weight decay of Adam optimizer as 0.1. For other hyper-parameters, we conduct a heuristic search by exploring the dropout rate from $\{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, the parameter ρ which controls the threshold of node reliability from $\{0.2, 0.4, 0.6, 0.8\}$ and the parameter γ which controls the proportion of knowledge transfer from $\{0.5, 1, 2, 5\}$ with the help of Optuna. Finally we select a set of hyper-parameters that make RDD perform best in the validation set.

(3) GraphAKD [10]: We train GraphAKD with same teacher GNN and student GNN. And we employ optimizer follows the original code. For other hyper-parameters, we conduct heuristic search by exploring the dropout rate from $\{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ for different GNNs, the parameter learning rate from $[0.001, 0.05]$, and the ratio k from $\{1, 5, 10, 20, 30\}$ with the help of optuna.

(4) FreeKD [5]: We fix FreeKD with learning rate as 0.05 and l_2 norm regularization weight decay as 0.0005 according to the original code. And we leverage optuna to explore the dropout rate from $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ and hidden dim from 32, 64, 128.

A.5 Hyper-parameters of FT

The detailed hyper-parameters used in FT are summarized in Table 5.

A.6 Hyper-parameters of LTD

For the hyper-parameter searching of LTD, we conduct heuristic search by exploring the learning rate of distillation step $\alpha \in [1e-7, 1e-2]$, the learning rate of learning step $\beta \in [1e-6, 1e-3]$ and the balance hyper-parameter λ from $\{0.1, 1, 50, 100, 200\}$ with

Table 8: Classification accuracies of different distillation frameworks on five GNN models.

Teacher	Dataset	GraphAKD	FreeKD	CPF	RDD	LTD
GCN	Cora	0.8515	0.8487	0.8585	0.8543	**0.8721**
	Citeseer	0.7320	0.7376	0.7552	0.7431	**0.7851**
	Pubmed	0.7974	0.8103	0.7842	0.8146	**0.8191**
	A-Computers	0.8540	0.8087	0.8644	0.8251	**0.8645**
	A-Photo	0.9304	0.9021	**0.9352**	0.8839	0.9324
GAT	Cora	0.8478	0.8403	0.8628	0.8464	**0.8656**
	Citeseer	0.7354	0.7370	0.7657	0.7481	**0.7735**
	Pubmed	0.8210	0.7993	0.7885	0.8218	**0.8274**
	A-Computers	0.8276	0.8275	0.8063	0.8006	**0.8304**
	A-Photo	0.9251	0.9111	0.9200	0.9112	**0.9316**
SAGE	Cora	0.8276	0.8506	0.8674	0.8567	**0.8703**
	Citeseer	0.7050	0.7536	0.7586	0.7470	**0.7746**
	Pubmed	0.7866	0.8138	0.8143	0.8173	**0.8401**
	A-Computers	0.7878	0.7996	0.7884	0.7986	**0.8144**
	A-Photo	0.8946	0.8626	0.8741	0.8084	**0.9306**
APPNP	Cora	0.8543	0.8656	0.8689	0.8642	**0.8693**
	Citeseer	0.7519	0.7503	0.7696	0.7580	**0.7851**
	Pubmed	0.8138	0.8291	0.8435	0.8387	**0.8436**
	A-Computers	0.8081	0.8087	0.8172	0.8112	**0.8363**
	A-Photo	0.9232	0.9246	**0.9337**	0.9255	**0.9337**
SGC	Cora	0.8468	0.8309	**0.8670**	0.8562	0.8660
	Citeseer	0.7547	0.7392	0.7713	0.7315	**0.7873**
	Pubmed	0.7937	0.8067	0.8205	0.8302	**0.8405**
	A-Computers	0.7853	0.8367	0.8023	0.8084	**0.8528**
	A-Photo	0.9145	0.9221	**0.9324**	0.9155	0.9297
Average ranking		4.04	3.84	2.48	3.36	1.12

the help of Optuna[5], an automatic hyper-parameter optimization toolkit. For other settings, we always use a 2-layer MLP in temperature parameterization with hidden dimension as 64 and dropout rate as 0.6.We restrict the temperatures within a reasonable range $[-0.2k, 0.8k]$ where $k = 1, 2, 3$. Note that we allow a negative temperature for distilling, which can help the student model correct the teacher's predictions more flexibly. For LTD$_+$, we also employ Optuna to explore balance hyper-parameters μ, ν, γ from $\{0.01, 0.1, 0.5, 1, 2, 3, 5, 10\}$. For all methods, we use early stopping with a patience of 50 and max epochs as 600. We have about 100 trials altogether, and finally select a set of hyper-parameters that make LTD perform best in the validation set. The detailed hyper-parameters used in LTD are summarized in Table 6.

A.7 Additional Experiments Analysis

A.7.1 Balance Hyper-parameter Analysis. We enumerate the balance hyper-parameter $\lambda \in \{0.1, 1, 50, 100, 200\}$, and evaluate the students distilled by LTD. Figure 7 shows the performance with respect to the balance hyper-parameter λ, and we can find that $\lambda = 1$ works well for all five GNN models. Since we use the same number of labeled nodes per class for all datasets, the best choice of λ is more related to the dataset: a larger dataset with less classes (*e.g.,* Pubmed) needs a larger λ to enhance the weight of ground truth instances. To summarize, λ is not sensitive to the choice of GNN models.

A.7.2 ***Performance under Different Training Ratios.*** Figure 8 presents the performance of LTD and FT under different training ratios. As the number of labeled nodes per class increases from 20 to 50, our proposed LTD consistently outperforms FT as well as the teacher GNN model by a large margin, which illustrates the robustness of LTD.

A.7.3 ***Generalization Gap Analysis.*** As we mentioned in Section 3.3.3, LTD can alleviate the overfitting issue compared with the traditional distillation framework. To support this statement, we report the performance of training/validation/test sets on A-Computers in Table 7. A-Computers has the largest number of classes and average node degree, and is more likely to cause the overfitting of message passing in our experiments. As we can see from the table, the students distilled by LTD have smaller generalization gaps between the performance on validation and test sets. Therefore, LTD can benefit from better generalization ability and outperform existing distillation frameworks.

A.7.4 ***Original Results of Figure 2.*** The original results of Figure 2 without being averaged are listed in Table 8. Our LTD has the best average ranking compared with SOTA distillation frameworks.

[5]https://optuna.org/