Table 6: 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

Table 7: 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

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

Teacher	Dataset	Teacher	CPF	RDD	LTD
GCN	Cora	0.8534	0.8585	0.8543	**0.8721**
	Citeseer	0.7359	0.7552	0.7431	**0.7851**
	Pubmed	0.7989	0.7842	0.8146	**0.8191**
	A-Computers	0.8594	0.8644	0.8251	**0.8645**
	A-Photo	0.9223	**0.9352**	0.8839	0.9324
GAT	Cora	0.8520	0.8628	0.8464	**0.8656**
	Citeseer	0.7525	0.7657	0.7481	**0.7735**
	Pubmed	0.7944	0.7885	0.8218	**0.8274**
	A-Computers	0.8091	0.8063	0.8006	**0.8304**
	A-Photo	0.9094	0.9200	0.9112	**0.9316**
GraphSAGE	Cora	0.8426	0.8674	0.8567	**0.8703**
	Citeseer	0.7276	0.7586	0.7470	**0.7746**
	Pubmed	0.8189	0.8143	0.8173	**0.8401**
	A-Computers	0.7829	0.7884	0.7986	**0.8144**
	A-Photo	0.9146	0.8741	0.8084	**0.9306**
APPNP	Cora	0.8581	0.8689	0.8642	**0.8693**
	Citeseer	0.7530	0.7696	0.7580	**0.7851**
	Pubmed	0.8301	0.8435	0.8387	**0.8436**
	A-Computers	0.8095	0.8172	0.8112	**0.8363**
	A-Photo	0.9225	**0.9337**	0.9255	**0.9337**
SGC	Cora	0.8454	**0.8670**	0.8562	0.8660
	Citeseer	0.7238	0.7713	0.7315	**0.7873**
	Pubmed	0.8205	0.8205	0.8302	**0.8405**
	A-Computers	0.8047	0.8023	0.8084	**0.8528**
	A-Photo	0.9118	**0.9324**	0.9155	0.9297
Average ranking		3.40	2.32	3.04	1.12

A DETAILS FOR REPRODUCIBILITY

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 follow the dataset settings in [17, 28], 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.3 Settings for Other Distillation Frameworks

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

- CPF [28]: 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 $\{0.0005, 0.001, 0.01\}$.

- RDD [31]: We fix RDD with learning rate as 0.1 and weight decay of Adam optimizer as 0.1. For other hyper-parameters, we conduct 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.

A.4 Hyper-parameters of FT

The detailed hyper-parameters used in FT are summarized in Table 6.

A.5 Hyper-parameters of LTD

In LTD, we use early stopping with max epochs as 600. 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. 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 7.

A.6 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.