1 Appendix: Supplementary Information

This document provides the supporting materials for our main paper, including the methodology for hyperparameter selection and detailed tables of inferred client label distribution.

1.1 Hyperparameters Selection

To determine the optimal values for the scalars a, b, and c for the loss function of the attack model, we conducted a grid search by evaluating the neural network loss across different parameter combinations. Each scalar was assigned candidate values of 0.0, 0.25, 0.5, and 1.0, and the combination yielding the lowest loss was selected as the optimal one. The optimal set of values for a, b, and c is provided in Table 1.

Dataset	**a**	**b**	**c**
Cora	0.000	0.500	0.500
PubMed	0.500	0.250	0.250
Citeseer	0.000	0.500	0.500
Amazon Computers	0.500	0.250	0.250

Table 1: Optimal values of the scalars a, b, and c in the loss function across the four datasets.

1.2 Label Distribution Inference

Tables 2, 3, 4, and 5 illustrate individual clients' inferred label distribution on the Cora, PubMed, Citeseer, and Amazon Computers datasets, respectively, when Graph Convolutional Network (GCN) was used for both FL training and shadow FL training. Similarly, Table 6, 7, 8, and 9 illustrate individual clients' inferred label distribution on the Cora, PubMed, Citeseer, and Amazon Computers datasets, respectively, when GraphSage was used for both FL training and shadow FL training. Table 10, 11, 12, and 13 illustrate individual clients' inferred label distribution on the Cora, PubMed, Citeseer, and Amazon Computers datasets, respectively, when Graph Isomorphism Network (GIN) was used for both FL training and shadow FL training.

Each table represents the result for 10 clients with varying label distributions, including equal, random, one class missing, single class only, and one class dominant distributions. The rows shaded in gray indicate the Ground Truth (GT) label proportions, while the unshaded rows show the attack model inferred distributions. Since our implementation provides label distribution estimates for all clients, any client can serve as the target for evaluating the attack's effectiveness under its specific distribution. Note that although the datasets differ in the number of classes, the number of clients is kept constant to have a consistent proportion setup across datasets.

Client	Label Proportion						
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
1	0.142	0.142	0.142	0.142	0.142	0.142	0.142
	0.158	0.100	0.114	0.221	0.179	0.137	0.087
2	0.128	0.123	0.143	0.282	0.153	0.138	0.030
	0.132	0.110	0.170	0.213	0.157	0.110	0.105
3	0.087	0.102	0.159	0.307	0.159	0.128	0.056
	0.136	0.101	0.132	0.221	0.166	0.101	0.141
4	0.123	0.071	0.174	0.312	0.169	0.082	0.066
	0.121	0.091	0.140	0.269	0.139	0.142	0.095
5	0.220	0.087	0.246	0.000	0.235	0.138	0.071
	0.097	0.060	0.356	0.206	0.103	0.103	0.072
6	0.133	0.076	0.164	0.282	0.164	0.128	0.051
	0.115	0.078	0.162	0.252	0.161	0.143	0.087
7	0.133	0.035	0.169	0.323	0.128	0.143	0.066
	0.097	0.060	0.356	0.206	0.103	0.103	0.072
8	0.000	0.000	1.000	0.000	0.000	0.000	0.000
	0.111	0.073	0.177	0.264	0.154	0.135	0.082
9	0.061	0.005	0.056	0.800	0.041	0.025	0.010
	0.101	0.078	0.149	0.401	0.136	0.073	0.059
10	0.061	0.005	0.056	0.800	0.041	0.025	0.010
	0.133	0.089	0.144	0.242	0.152	0.146	0.091

Table 2: Table illustrating the label proportions across clients in the Cora dataset. The proportions were derived using a GCN for both FL training and shadow FL training, while the attack model employed was a neural-network-based approach. Rows shaded with gray color are the ground truth label distribution, and unshaded rows are the inferred distribution.

Client	Label Proportion		
	Class 1	Class 2	Class 3
1	0.333	0.333	0.333
	0.304	0.397	0.298
2	0.200	0.393	0.406
	0.247	0.366	0.386
3	0.191	0.404	0.403
	0.252	0.352	0.395
4	0.207	0.395	0.397
	0.250	0.338	0.410
5	0.207	0.389	0.403
	0.250	0.340	0.409
6	0.216	0.385	0.397
	0.309	0.389	0.300
7	0.000	0.372	0.415
	0.246	0.366	0.386
8	0.000	0.000	1.000
	0.245	0.377	0.377
9	0.071	0.128	0.799
	0.250	0.351	0.399
10	0.071	0.128	0.799
	0.247	0.365	0.387

Table 3: Table illustrating the label proportions across clients in the PubMed dataset. The proportions were derived using a GCN for both FL training and shadow FL training, while the attack model employed was a neural-network-based approach. Rows shaded with gray color are the ground truth label distribution, and unshaded rows are the inferred distribution.

Client	**Label Proportion**					
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
1	0.166	0.166	0.166	0.166	0.166	0.166
	0.108	0.135	0.170	0.220	0.205	0.159
2	0.062	0.200	0.208	0.204	0.179	0.145
	0.113	0.172	0.186	0.200	0.178	0.148
3	0.079	0.183	0.225	0.187	0.195	0.129
	0.096	0.150	0.180	0.227	0.187	0.158
4	0.091	0.141	0.195	0.208	0.166	0.195
	0.107	0.207	0.170	0.196	0.170	0.147
5	0.100	0.242	0.200	0.000	0.230	0.225
	0.107	0.186	0.164	0.211	0.183	0.146
6	0.070	0.158	0.183	0.258	0.162	0.166
	0.100	0.126	0.189	0.238	0.185	0.158
7	0.079	0.175	0.187	0.200	0.195	0.162
	0.111	0.166	0.239	0.154	0.165	0.162
8	0.000	0.000	1.000	0.000	0.000	0.000
	0.111	0.166	0.239	0.154	0.165	0.162
9	0.012	0.029	0.071	0.799	0.046	0.041
	0.161	0.152	0.176	0.194	0.166	0.148
10	0.012	0.029	0.071	0.799	0.046	0.041
	0.122	0.157	0.181	0.203	0.182	0.153

Table 4: Table illustrating the label proportions across clients in the Citeseer dataset. The proportions were derived using a GCN for both FL training and shadow FL training, while the attack model employed was a neural-network-based approach. Rows shaded with gray color are the ground truth label distribution, and unshaded rows are the inferred distribution.

Client	Label Proportion									
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10
1	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100
	0.031	0.160	0.078	0.034	0.377	0.024	0.032	0.047	0.196	0.0178
2	0.033	0.162	0.104	0.047	0.366	0.024	0.041	0.049	0.152	0.018
	0.031	0.156	0.075	0.031	0.374	0.024	0.031	0.047	0.209	0.018
3	0.030	0.160	0.100	0.038	0.375	0.032	0.028	0.066	0.150	0.017
	0.036	0.160	0.080	0.032	0.303	0.026	0.035	0.057	0.243	0.023
4	0.032	0.168	0.126	0.000	0.384	0.020	0.031	0.060	0.155	0.020
	0.031	0.157	0.077	0.031	0.358	0.026	0.031	0.048	0.218	0.0189
5	0.043	0.154	0.094	0.039	0.378	0.023	0.037	0.060	0.154	0.014
	0.032	0.153	0.076	0.035	0.367	0.024	0.032	0.044	0.216	0.018
6	0.036	0.155	0.111	0.041	0.381	0.015	0.036	0.056	0.142	0.024
	0.035	0.161	0.078	0.036	0.361	0.026	0.034	0.050	0.196	0.020
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
	0.064	0.154	0.098	0.065	0.230	0.070	0.063	0.075	0.130	0.046
8	0.035	0.162	0.119	0.037	0.352	0.020	0.034	0.057	0.150	0.031
	0.047	0.117	0.080	0.053	0.300	0.035	0.054	0.056	0.204	0.0503
9	0.012	0.045	0.025	0.009	0.800	0.009	0.009	0.025	0.059	0.005
	0.048	0.147	0.122	0.139	0.110	0.051	0.136	0.069	0.113	0.059
10	0.012	0.045	0.025	0.009	0.800	0.009	0.009	0.025	0.059	0.005
	0.047	0.117	0.080	0.053	0.300	0.035	0.054	0.056	0.204	0.050

Table 5: Table illustrating the label proportions across clients in the Amazon Computers dataset. The proportions were derived using a GCN for both FL training and shadow FL training, while the attack model employed was a neural-network-based approach. Rows shaded with gray color are the ground truth label distribution, and unshaded rows are the inferred distribution.

Client	Label Proportion						
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
1	0.142	0.142	0.142	0.142	0.142	0.142	0.142
	0.098	0.015	0.412	0.201	0.110	0.096	0.064
2	0.128	0.123	0.143	0.282	0.153	0.138	0.030
	0.142	0.038	0.138	0.243	0.158	0.151	0.127
3	0.087	0.102	0.159	0.307	0.159	0.128	0.056
	0.098	0.017	0.423	0.190	0.109	0.097	0.064
4	0.123	0.071	0.174	0.312	0.169	0.082	0.066
	0.106	0.020	0.377	0.196	0.118	0.107	0.072
5	0.220	0.087	0.246	0.000	0.235	0.138	0.071
	0.020	0.001	0.879	0.053	0.020	0.017	0.008
6	0.133	0.076	0.164	0.282	0.164	0.128	0.051
	0.136	0.043	0.190	0.212	0.152	0.144	0.119
7	0.133	0.035	0.169	0.323	0.128	0.143	0.066
	0.063	0.011	0.638	0.126	0.067	0.055	0.037
8	0.000	0.000	1.000	0.000	0.000	0.000	0.000
	0.114	0.019	0.310	0.225	0.131	0.117	0.081
9	0.061	0.005	0.056	0.800	0.041	0.025	0.010
	0.096	0.013	0.420	0.206	0.108	0.093	0.060
10	0.061	0.005	0.056	0.800	0.041	0.025	0.010
	0.020	0.001	0.879	0.053	0.020	0.017	0.008

Table 6: Table illustrating the label proportions across clients in the Cora dataset. The proportions were derived using a GraphSage for both FL training and shadow FL training, while the attack model employed was a neural-network-based approach. Rows shaded with gray color are the ground truth label distribution, and unshaded rows are the inferred distribution.

Client	Label Proportion		
	Class 1	Class 2	Class 3
1	0.333	0.333	0.333
	0.097	0.186	0.716
2	0.200	0.393	0.406
	0.096	0.184	0.718
3	0.191	0.404	0.403
	0.096	0.177	0.726
4	0.207	0.395	0.397
	0.100	0.191	0.708
5	0.207	0.389	0.403
	0.097	0.186	0.716
6	0.216	0.385	0.397
	0.096	0.177	0.726
7	0.000	0.372	0.415
	0.094	0.169	0.736
8	0.000	0.000	1.000
	0.091	0.173	0.734
9	0.071	0.128	0.799
	0.093	0.179	0.726
10	0.071	0.128	0.799
	0.098	0.185	0.715

Table 7: Table illustrating the label proportions across clients in the PubMed dataset. The proportions were derived using a GraphSage for both FL training and shadow FL training, while the attack model employed was a neural-network-based approach. Rows shaded with gray color are the ground truth label distribution, and unshaded rows are the inferred distribution.

Client	Label Proportion					
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
1	0.166	0.166	0.166	0.166	0.166	0.166
	0.062	0.099	0.632	0.020	0.091	0.093
2	0.062	0.200	0.208	0.204	0.179	0.145
	0.125	0.184	0.276	0.019	0.216	0.177
3	0.079	0.183	0.225	0.187	0.195	0.129
	0.079	0.122	0.539	0.025	0.117	0.115
4	0.091	0.141	0.195	0.208	0.166	0.195
	0.013	0.033	0.894	0.006	0.025	0.027
5	0.100	0.242	0.200	0.000	0.230	0.225
	0.050	0.091	0.676	0.031	0.071	0.077
6	0.070	0.158	0.183	0.258	0.162	0.166
	0.050	0.091	0.676	0.031	0.071	0.077
7	0.079	0.175	0.187	0.200	0.195	0.162
	0.053	0.087	0.679	0.017	0.081	0.080
8	0.000	0.000	1.000	0.000	0.000	0.000
	0.131	0.174	0.298	0.036	0.190	0.169
9	0.012	0.029	0.071	0.799	0.046	0.041
	0.099	0.146	0.447	0.027	0.142	0.136
10	0.012	0.029	0.071	0.799	0.046	0.041
	0.024	0.047	0.837	0.010	0.040	0.040

Table 8: Table illustrating the label proportions across clients in the Citeseer dataset. The proportions were derived using a GraphSage for both FL training and shadow FL training, while the attack model employed was a neural-network-based approach. Rows shaded with gray color are the ground truth label distribution, and unshaded rows are the inferred distribution.

Client	Label Proportion									
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10
1	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100
	0.052	0.050	0.096	0.157	0.169	0.059	0.065	0.175	0.104	0.069
2	0.033	0.162	0.104	0.047	0.366	0.024	0.041	0.049	0.152	0.018
	0.080	0.070	0.104	0.104	0.150	0.083	0.090	0.108	0.114	0.094
3	0.030	0.160	0.100	0.038	0.375	0.032	0.028	0.066	0.150	0.017
	0.055	0.050	0.114	0.109	0.216	0.064	0.081	0.089	0.141	0.077
4	0.032	0.168	0.126	0.000	0.384	0.020	0.031	0.060	0.155	0.020
	0.055	0.054	0.084	0.218	0.125	0.063	0.063	0.181	0.086	0.066
5	0.043	0.154	0.094	0.039	0.378	0.023	0.037	0.060	0.154	0.014
	0.063	0.056	0.114	0.094	0.194	0.068	0.086	0.102	0.133	0.085
6	0.036	0.155	0.111	0.041	0.381	0.015	0.036	0.056	0.142	0.024
	0.042	0.038	0.059	0.156	0.093	0.042	0.047	0.409	0.061	0.048
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
	0.057	0.051	0.070	0.221	0.098	0.061	0.058	0.250	0.070	0.059
8	0.035	0.162	0.119	0.037	0.352	0.020	0.034	0.057	0.150	0.031
	0.059	0.054	0.079	0.119	0.127	0.056	0.061	0.284	0.086	0.070
9	0.012	0.045	0.025	0.009	0.800	0.009	0.009	0.025	0.059	0.005
	0.060	0.054	0.115	0.097	0.205	0.067	0.085	0.091	0.139	0.082
10	0.012	0.045	0.025	0.009	0.800	0.009	0.009	0.025	0.059	0.005
	0.006	0.007	0.006	0.188	0.008	0.005	0.004	0.763	0.006	0.005

Table 9: Table illustrating the label proportions across clients in the Amazon Computers dataset. The proportions were derived using a GraphSage for both FL training and shadow FL training, while the attack model employed was a neural-network-based approach. Rows shaded with gray color are the ground truth label distribution, and unshaded rows are the inferred distribution.

Client	Label Proportion						
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
1	0.142	0.142	0.142	0.142	0.142	0.142	0.142
	0.134	0.109	0.154	0.237	0.169	0.121	0.074
2	0.128	0.123	0.143	0.282	0.153	0.138	0.030
	0.136	0.106	0.152	0.234	0.171	0.124	0.075
3	0.087	0.102	0.159	0.307	0.159	0.128	0.056
	0.134	0.108	0.154	0.237	0.169	0.121	0.074
4	0.123	0.071	0.174	0.312	0.169	0.082	0.066
	0.133	0.109	0.154	0.237	0.169	0.121	0.074
5	0.220	0.087	0.246	0.000	0.235	0.138	0.071
	0.135	0.108	0.153	0.239	0.168	0.120	0.073
6	0.133	0.076	0.164	0.282	0.164	0.128	0.051
	0.135	0.112	0.156	0.231	0.167	0.118	0.078
7	0.133	0.035	0.169	0.323	0.128	0.143	0.066
	0.134	0.107	0.156	0.237	0.169	0.120	0.073
8	0.000	0.000	1.000	0.000	0.000	0.000	0.000
	0.133	0.107	0.154	0.235	0.170	0.122	0.075
9	0.061	0.005	0.056	0.800	0.041	0.025	0.010
	0.132	0.110	0.157	0.238	0.166	0.119	0.074
10	0.061	0.005	0.056	0.800	0.041	0.025	0.010
	0.136	0.106	0.152	0.234	0.171	0.124	0.075

Table 10: Table illustrating the label proportions across clients in the Cora dataset. The proportions were derived using a GIN for both FL training and shadow FL training, while the attack model employed was a neural-network-based approach. Rows shaded with gray color are the ground truth label distribution, and unshaded rows are the inferred distribution.

Client	Label Proportion		
	Class 1	Class 2	Class 3
1	0.333	0.333	0.333
	0.232	0.377	0.389
2	0.200	0.393	0.406
	0.225	0.379	0.395
3	0.191	0.404	0.403
	0.225	0.379	0.395
4	0.207	0.395	0.397
	0.230	0.379	0.390
5	0.207	0.389	0.403
	0.225	0.380	0.394
6	0.216	0.385	0.397
	0.231	0.377	0.390
7	0.000	0.372	0.415
	0.229	0.383	0.387
8	0.000	0.000	1.000
	0.234	0.374	0.391
9	0.071	0.128	0.799
	0.232	0.377	0.389
10	0.071	0.128	0.799
	0.234	0.375	0.390

Table 11: Table illustrating the label proportions across clients in the PubMed dataset. The proportions were derived using a GIN for both FL training and shadow FL training, while the attack model employed was a neural-network-based approach. Rows shaded with gray color are the ground truth label distribution, and unshaded rows are the inferred distribution.

Client	Label Proportion					
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
1	0.166	0.166	0.166	0.166	0.166	0.166
	0.114	0.171	0.197	0.190	0.180	0.146
2	0.062	0.200	0.208	0.204	0.179	0.145
	0.114	0.169	0.196	0.190	0.181	0.147
3	0.079	0.183	0.225	0.187	0.195	0.129
	0.114	0.171	0.198	0.188	0.181	0.146
4	0.091	0.141	0.195	0.208	0.166	0.195
	0.118	0.174	0.196	0.187	0.177	0.145
5	0.100	0.242	0.200	0.000	0.230	0.225
	0.115	0.171	0.196	0.187	0.183	0.146
6	0.070	0.158	0.183	0.258	0.162	0.166
	0.115	0.172	0.195	0.188	0.180	0.147
7	0.079	0.175	0.187	0.200	0.195	0.162
	0.115	0.171	0.196	0.187	0.183	0.146
8	0.000	0.000	1.000	0.000	0.000	0.000
	0.114	0.172	0.197	0.188	0.179	0.146
9	0.012	0.029	0.071	0.799	0.046	0.041
	0.117	0.168	0.201	0.184	0.184	0.144
10	0.012	0.029	0.071	0.799	0.046	0.041
	0.115	0.173	0.200	0.187	0.180	0.143

Table 12: Table illustrating the label proportions across clients in the Citeseer dataset. The proportions were derived using a GIN for both FL training and shadow FL training, while the attack model employed was a neural-network-based approach. Rows shaded with gray color are the ground truth label distribution, and unshaded rows are the inferred distribution.

Client	Label Proportion									
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10
1	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100
	0.031	0.024	0.350	0.161	0.158	0.040	0.037	0.049	0.109	0.037
2	0.033	0.162	0.104	0.047	0.366	0.024	0.041	0.049	0.152	0.018
	0.016	0.015	0.529	0.213	0.087	0.021	0.019	0.024	0.055	0.018
3	0.030	0.160	0.100	0.038	0.375	0.032	0.028	0.066	0.150	0.017
	0.052	0.039	0.165	0.111	0.193	0.076	0.071	0.080	0.131	0.080
4	0.032	0.168	0.126	0.000	0.384	0.020	0.031	0.060	0.155	0.020
	0.045	0.033	0.208	0.1323	0.192	0.066	0.059	0.071	0.126	0.066
5	0.043	0.154	0.094	0.039	0.378	0.023	0.037	0.060	0.154	0.014
	0.053	0.040	0.159	0.098	0.204	0.078	0.070	0.082	0.132	0.081
6	0.036	0.155	0.111	0.041	0.381	0.015	0.036	0.056	0.142	0.024
	0.034	0.024	0.185	0.098	0.283	0.055	0.050	0.066	0.144	0.058
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
	0.006	0.007	0.785	0.088	0.052	0.007	0.007	0.010	0.028	0.006
8	0.035	0.162	0.119	0.037	0.352	0.020	0.034	0.057	0.150	0.031
	0.045	0.033	0.242	0.135	0.176	0.056	0.051	0.068	0.131	0.057
9	0.012	0.045	0.025	0.009	0.800	0.009	0.009	0.025	0.059	0.005
	0.007	0.006	0.738	0.144	0.047	0.007	0.006	0.009	0.029	0.006
10	0.012	0.045	0.025	0.009	0.800	0.009	0.009	0.025	0.059	0.005
	0.042	0.033	0.265	0.153	0.162	0.059	0.053	0.063	0.109	0.058

Table 13: Table illustrating the label proportions across clients in the Amazon Computers dataset. The proportions were derived using a GIN for both FL training and shadow FL training, while the attack model employed was a neural-network-based approach. Rows shaded with gray color are the ground truth label distribution, and unshaded rows are the inferred distribution.