A Details for Reproducibility

In this section, we offer comprehensive details to facilitate reproducibility.

A.1 Experimental Environments

We implement our method using PyTorch Geometric. We have also listed the required Python environment in the *requirement.txt* and all the commands used for the experiments in the *experiment.sh* with the code we have submitted. All our experiments are run on a single GPU device of GeForce GTX 3090 with 22 GB memory, and the operating system is Red Hat 7.3.1-5.

Table 3: The statistics of German and Credit. Here "hete-" is short for "heterogeneous".

Dataset	German	Credit
# Nodes	1,000	30,000
# Features	27	13
# Edges	22,242	1,436,858
Node label	Credit status	Future default
Sensitive attribute	Gender	Region
Avg. degree	44.48	95.79
Avg. hete-degree	8.48	3.83
# Nodes w/o hete-neighbors	30	9,697

A.2 Dataset Details

Here, we provide a detailed description of all five fairness-oriented datasets that are used to validate our proposed FairSIN, as follows:

German Credit (German): The German Credit dataset is primarily utilized for predicting the likelihood of loan default, where clients are represented as nodes, and edges are formed based on similarities in their credit accounts. The dataset contains various features related to creditworthiness, such as credit history, loan duration, credit amount, employment status, personal status, age, and more. The classification task involves determining whether a client's credit risk is high or low while taking into account the sensitive attribute of gender.

Credit Defaulter (Credit): The Credit Defaulter is a widely used dataset in research, which contains information about credit default of customers in a bank. This dataset includes various features such as demographic information, credit history, and other factors that may affect credit default, such as outstanding balance, payment history, and age of the customer. In this scenario, the nodes represent credit card users, and edges connect users who exhibit similar patterns in purchases and payments. The objective is to predict whether a user will default on their credit card payment, with age serving as the sensitive attribute.

Recidivism (Bail): The Recidivism dataset comprises information on individuals who were granted bail between 1990-2009, including their demographic details. In this dataset, nodes represent defendants released on bail during the stated period, and edges connect defendants based on their shared past criminal records and demographics. The objective is to predict whether a defendant is more likely to commit a violent or non-violent crime upon their release, while also considering the sensitive attribute of race.

Pokec: Pokec-n and Pokec-z are two datasets sampled from the large Slovakian social network, Pokec. This social network is known to be the most popular in Slovakia and includes user features such as gender, age, hobbies, interests, education, and working field, among others. In these datasets, nodes represent users on the Pokec social network, and edges connect users who share a social connection. The objective is to predict the working field of the users while considering the sensitive attribute of region.

Each dataset is partitioned into 50%/25%/25% for training, validation, and testing respectively. The statistical information for the German and Credit datasets is provided in Table 3, while the statistics for the remaining three datasets have been previously presented in Table 1.

A.3 Implementation details

We implement all baseline methods using the source code provided by the authors, and train on the five datasets with the same pre-processing and dataset partition. To ensure a fair comparison, all hyperparameters are tuned based on the guidelines provided in the author's paper or Github repository, and the best results are reported. Following previous work (Wang et al. 2022c; Dai and Wang 2021) , we use the validation set to find the best trade-off, i.e. we use the trade-off value computed in each epoch on the validation set as the criterion for selecting parameters, which is then used to evaluate the test set. For each baseline, we provide the parameter tuning space in detail as follows:

FairGNN: We use a dropout rate of $\{0.0, 0.5, 0.8\}$, 32 hidden layers, and learning rates of $\{0.0001, 0.001, 0.01\}$. Additionally, we set the sensitive number to 200 and the label number to 3,000.

EDITS: We set the initial learning rate to 0.003, and the threshold proportions for the Credit, German, and Recidivism datasets are set to $\{0.02, 0.25, 0.012\}$, respectively.

NIFTY: We set the dropout rate to $\{0.0, 0.5, 0.8\}$, the number of hidden layers to 16, and the learning rate to $\{0.01, 0.001 , 0.0001\}$. Additionally, we used a drop edge rate of 0.001, drop feature rate of 0.1, and regularization coefficients of 0.6."

FairVGNN: We use a dropout rate of $\{0.0, 0.5, 0.8\}$, and $K = 10$. For each module, we search for the learning rate in the range of $\{0.001, 0.01\}$, and set the training epochs of each iteration to $\{5, 10\}$. Additionally, we use regularization coefficients of $\alpha \in \{0,0.5,1\}$, and set the prefix cutting threshold to $\epsilon \in \{0.01,0.1,1\}$.

For parameters not mentioned above, we use the default setting.

B Theoretical Analysis

In this section, we will provide a proof for the Theorem 1 presented in the main text and complete the relevant definitions for $\hat{P}_\theta$.

Table 4: Comparison between FairSIN and SOTA methods on German and Credit. (**Bold**: the best; underline: the runner-up.)

Encoder	Method	German				Credit			
		F1↑	ACC↑	DP↓	EO↓	F1↑	ACC↑	DP↓	EO↓
GCN	vanilla	81.63±0.68	**72.28±1.52**	32.43±10.29	24.69±7.74	82.32±0.03	74.13±0.04	12.44±0.06	10.24±0.09
	FairGNN	82.01±0.26	69.68±0.30	3.49±2.15	3.40±2.15	81.84±1.19	73.41±1.24	12.64±2.11	10.41±2.03
	EDITS	81.55±0.59	71.60±0.89	4.05±4.48	3.89±4.23	81.81±0.28	73.51±0.30	10.90±1.22	8.75±1.21
	NIFTY	81.40±0.54	69.92±1.14	5.73±5.25	5.08±4.29	81.72±0.05	73.45±0.06	11.68±0.07	9.39±0.07
	FairVGNN	82.14±0.42	70.16±0.86	1.71±1.68	0.88±0.58	87.08±0.74	**78.04±0.33**	5.02±5.22	3.60±4.31
	FairSIN	**82.34±0.01**	70.08±0.16	**0.22±0.43**	**0.02±0.04**	**87.56±0.01**	77.87±0.01	**0.50±0.70**	**0.25±0.34**
GIN	vanilla	82.14±0.86	**72.96±1.14**	13.94±6.81	9.08±6.04	85.42±1.15	77.39±1.00	5.66±1.82	3.47±1.72
	FairGNN	**83.16±0.56**	72.24±1.44	6.88±4.42	2.06±1.46	79.47±5.29	70.33±5.50	4.67±3.06	3.94±1.49
	EDITS	82.80±0.22	72.08±0.66	0.86±0.76	1.72±1.14	82.47±0.85	74.07±0.98	14.11±14.45	15.40±15.76
	NIFTY	80.46±3.06	69.92±3.64	5.26±3.24	5.34±5.67	84.05±0.82	75.59±0.66	7.09±4.62	6.22±3.26
	FairVGNN	82.40±0.14	70.16±0.32	0.43±0.54	0.34±0.41	87.44±0.23	**78.18±0.20**	2.85±2.01	1.72±1.80
	FairSIN	82.52±0.33	70.40±0.80	**0.30±0.29**	**0.19±0.33**	**87.56±0.01**	77.88±0.12	**0.36±0.72**	**0.23±0.45**
SAGE	vanilla	81.25±1.72	**72.12±1.76**	20.33±11.82	14.86±10.96	85.06±1.64	76.77±0.68	14.31±6.55	11.78±5.71
	FairGNN	82.29±0.32	70.64±0.74	7.65±8.07	4.18±4.86	83.97±2.00	75.29±1.62	6.17±5.57	5.06±4.46
	EDITS	81.04±1.09	71.68±1.25	8.42±7.35	5.69±2.16	82.41±0.52	74.13±0.59	11.34±6.36	9.38±5.39
	NIFTY	79.20±1.19	69.60±1.50	7.74±7.80	5.17±2.38	82.60±1.25	74.39±1.35	10.65±1.65	8.10±1.91
	FairVGNN	81.91±0.63	70.00±0.25	1.36±1.90	1.22±1.49	87.84±0.32	**79.94±0.30**	4.94±1.10	2.39±0.71
	FairSIN	**82.53±0.27**	70.40±0.62	**0.32±0.25**	**0.08±0.33**	**87.84±0.23**	78.91±0.61	**1.38±1.71**	**0.79±0.94**

B.1 Proof for Theorem 1

Here, we present a proof for Theorem 1, illustrating how message passing can result in an elevation of bias within the context of sensitive information neutralization. To begin, let's revisit Theorem 1.

Theorem 1. *Assume that node representations are biased and can be identified by the predictor, i.e., $\mu_c > \mu_{ic}$. For node v_i, we consider a message passing process that updates x_i by $x_i' = x_i + x_i^{neigh}$. Then we have*

$$\mathbb{E}\{\mathcal{D}_\theta(s_i|x_i') - \mathcal{D}_\theta(\bar{s}_i|x_i')\} > \mathbb{E}\{\mathcal{D}_\theta(s_i|x_i) - \mathcal{D}_\theta(\bar{s}_i|x_i)\}, \tag{7}$$

which means that the predictor $\hat{P}_\theta$ can identify the sensitive attributes more accurately.

Proof. *Recall that $\mathcal{D}_\theta$ is a linear function, and thus we have*

$$\begin{aligned}
&\mathbb{E}\{\mathcal{D}_\theta(s_i|x_i') - \mathcal{D}_\theta(\bar{s}_i|x_i')\} \\
=&\mathbb{E}\{\mathcal{D}_\theta(s_i|x_i) + \mathcal{D}_\theta(s_i|x_i^{neigh}) - \mathcal{D}_\theta(\bar{s}_i|x_i) - \mathcal{D}_\theta(\bar{s}_i|x_i^{neigh})\} \\
=&\mathbb{E}\{\mathcal{D}_\theta(s_i|x_i) - \mathcal{D}_\theta(\bar{s}_i|x_i)\} + P_i^{same}\mathbb{E}\{\mathcal{D}_\theta(s_i|x_i^{same}) - \\
&\mathcal{D}_\theta(\bar{s}_i|x_i^{same})\} - P_i^{diff}\mathbb{E}\{\mathcal{D}_\theta(\bar{s}_i|x_i^{diff}) - \mathcal{D}_\theta(s_i|x_i^{diff})\} \\
=&\mathbb{E}\{\mathcal{D}_\theta(s_i|x_i) - \mathcal{D}_\theta(\bar{s}_i|x_i)\} + (P_i^{same} - P_i^{diff})(\mu_c - \mu_{ic}) \\
>&\mathbb{E}\{\mathcal{D}_\theta(s_i|x_i) - \mathcal{D}_\theta(\bar{s}_i|x_i)\}.
\end{aligned} \tag{8}$$

B.2 The Design of Intensity Function

Biases are quantified using the mean $\hat{P}_\theta(s|x)$, with higher values indicating a more significant susceptibility to sensitive leakage within the representations. Here, we establish the parameterized predictor $\hat{P}_\theta(s|x)$ by normalizing the intensity function $\mathcal{D}_\theta$, which in turn is employed to compute $\hat{P}_\theta(s|x)$:

$$\begin{aligned}
\hat{P}_\theta(s|x) &= \frac{1}{2} + \frac{1}{2}\tanh(\mathcal{D}_\theta(s|x) - \mathcal{D}_\theta(\bar{s}|x)), \\
\hat{P}_\theta(\bar{s}|x) &= \frac{1}{2} + \frac{1}{2}\tanh(\mathcal{D}_\theta(\bar{s}|x) - \mathcal{D}_\theta(s|x)).
\end{aligned} \tag{9}$$

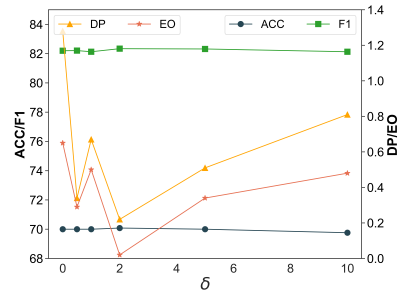
Be aware that we did not use the conditional entropy since the log operation has numerical instability issues in practice. Other normalization functions may also exist.

C Additional Experiment Analysis

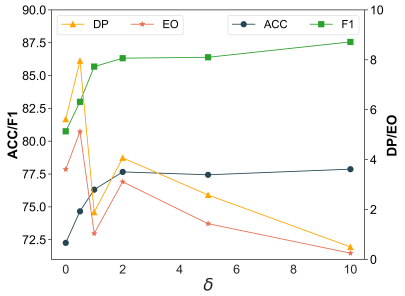
In this section, we will present experimental outcomes for the datasets not fully covered in the main text due to space limitations.

C.1 Effectiveness of FairSIN on German and Credit

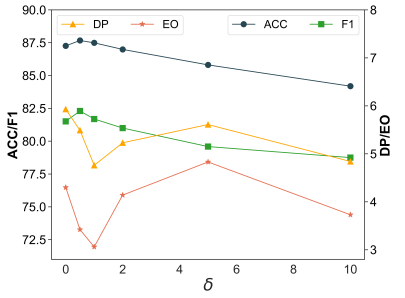
Here we present the outcomes of FairSIN applied to the German and Credit datasets. By leveraging our neutralization strategy, we also attain a better trade-off compared to state-of-the-art (SOTA) techniques. As displayed in Table 4, FairSIN demonstrates the finest overall classification performance and group fairness. Additionally, the Credit dataset exhibit limited heterogeneous neighbors, as indicated in Table 3, thus achieving a significant fairness improvement.



(a) German



(b) Credit



(c) Bail

Figure 6: Classification performance and group fairness under different values of hyper-parameter δ.

C.2 Hyper-parameter Analysis on German, Credit and Bail

For the remaining three datasets, namely German, Credit, and Bail, we perform a hyperparameter analysis and present the results shown in Figure 6. Also, we explore the impact of the hyperparameter δ across values $\{0, 0.5, 1, 2, 5, 10\}$ using a GCN encoder. Regarding the German and Bail datasets, we observe that an optimal value of $\delta = 2$ or $\delta = 1$ results in a favorable balance between predictive performance and fairness. Notably, in the case of the Bail dataset, which possesses the most heterogeneous neighbors, a smaller δ is necessary. Conversely, due to the sparsity of information from heterogeneous neighbors in the Credit dataset, both predictive performance and fairness improve when δ is increased to 10. These findings align closely with our concept of *neutralization*.

C.3 Sensitive Biases Analysis on German and Credit

We conducted an additional experiment to analyze bias changes on the German and Credit dataset during message passing and neutralization as shown in Figure 7. Similar to Figure 3, message passing exacerbates the sensitive bias, while neutralization has the opposite effect. These results validate our theoretical analysis and demonstrate that the advantage of our neutralization-based methods.

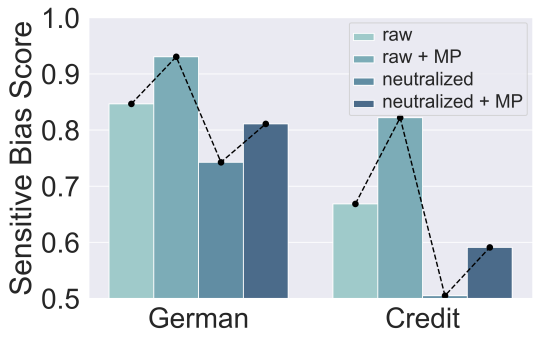


Figure 7: Sensitive biases in four groups of features on German and Credit. The biases are measured by average $\hat{P}_\theta(s|x)$, and larger scores indicate more serious sensitive leakage in the representations.