Towards a Unified Framework for Fair and Stable Graph Representation Learning: Supplementary material

Chirag Agarwal[1] **Himabindu Lakkaraju***[1] **Marinka Zitnik***[1]

[1]Harvard University

A PROPOSITION 1 AND ITS PROOF

Proposition 1 (Counterfactual Fairness of Downstream Classifier). *If the representations learned by our framework* NIFTY *satisfy counterfactual fairness, then a downstream classifier $f : \mathbf{z}_u \rightarrow \hat{y}_u$ which leverages these representations also satisfies counterfactual fairness.*

Proof. The downstream classifier uses the representation $\mathbf{z}_u$ output by our framework for predicting the label $\hat{y}_u$ of node u, thus forming a Markov chain $\mathbf{x}_u \rightarrow \mathbf{z}_u \rightarrow \hat{y}_u$ [Liao et al., 2019]. As we discuss in Section 3, node representations are said to be counterfactually fair if they are independent of the sensitive attribute. *i.e.,* the mutual information between the sensitive attribute s and the representation $\mathbf{z}_u$ for any given node u is zero: $I(s; \mathbf{z}_u) = 0$.

Using the properties of inequality and non-negativity of mutual information:

$$0 \leq I(s; \hat{y}_u) \leq I(s; \mathbf{z}_u) \text{ and } I(s; \mathbf{z}_u){=}0 \implies I(s; \hat{y}_u){=}0 \tag{1}$$

Therefore, the node label $\hat{y}_u$ for any given node u is independent of the sensitive attribute s, and consequently the downstream node classifier satisfies counterfactual fairness.

B DATASET DETAILS

German Credit Graph. The German Graph credit dataset classifies people described by a set of attributes as good or bad credit risks [Dua and Graff, 2017]. It consists of attributes like Gender, LoanAmount, and other account-related features of 1,000 clients. We use Minkowski distance as the similarity measure for calculating the similarity between two node attributes using: $1/(1 + \text{minkowski}(\mathbf{x}_u, \mathbf{x}_v))$. To obtain the credit graph network that connects clients, we connect two nodes if the similarity between them is 80% of the maximum similarity between all respective nodes (Refer Table. 1 for details). We argue that a graph neural network is fair if it predicts the client credit risk irrespective of their gender. Hence, we used *gender* as the sensitive attribute for the loan dataset.

Recidivism Graph. The dataset consists of samples of bail outcomes collected from several state courts in the US between 1990-2009 [Jordan and Freiburger, 2015]. It consists of past criminal records, demographic attributes, and other details of 18,876 defendants who got released on bail. We use Minkowski distance as the similarity measure for calculating the similarity between two node attributes using: $1/(1 + \text{minkowski}(\mathbf{x}_u, \mathbf{x}_v))$. To obtain the bail graph network that connects defendants, we connect two nodes if the similarity between them is 60% of the maximum similarity between all respective nodes (Refer Table. 1 for details). A machine learning model is trained to predict a defendant who is more likely to commit a violent or nonviolent crime once released on bail. A fair model should make predictions independent of the defendant's *race*, and, thus, we use it as the protected attribute for the dataset.

*Equal Contribution

Credit Defaulter Graph. We use a processed version [Ustun et al., 2019] of the credit dataset in Yeh and Lien [2009]. The task is to predict whether an applicant will default on an upcoming credit card payment. The dataset contains 30,000 individuals with features like education, credit history, age, and features derived from their spending and payment patterns. We use Minkowski distance as the similarity measure for calculating the similarity between two node attributes using: $1/(1 + \text{minkowski}(\mathbf{x}_u, \mathbf{x}_v))$. To obtain the credit defaulter graph network that connects applicants, we connect two nodes if the similarity between them is 70% of the maximum similarity between all respective nodes (Refer Table. 1 for details). For the credit dataset, we used *age* as the sensitive attribute.

Table 1: Statistics of novel graph datasets designed for node classification and accompanied by sensitive attributes. The datasets are appropriate to study fairness- and stability-aware algorithms.

Dataset	**German credit graph**	**Recidivism graph**	**Credit defaulter graph**
Nodes	1,000	18,876	30,000
Edges	22,242	321,308	1,436,858
Node features	27	18	13
Average node degree	44.48±26.51	34.04±46.65	95.79±85.88
Sensitive attribute	Gender (Male/Female)	Race (Black/White)	Age ($\leq 25/ > 25$)
Node labels	good credit vs. bad credit	bail vs. no bail	payment default vs. no default

C ARCHITECTURE AND HYPERPARAMETER SELECTION

We provide an overview of the important components of our proposed architecture and their respective training settings.

Encoder. The encoder block of our proposed framework can comprise of either simple Multilayer Perceptron (MLP) networks or any other GNN variant. For all our experiments, we use the vanilla GNN as the encoder block of our contrastive learning framework. For all datasets, we use a single-layer GNN encoder and set the hidden dimensionality to 16. The encoder is followed by a two-layer MLP projection head [Chen et al., 2020]. We only use ReLU and BatchNormalization (BN) layers after the first hidden layer in the MLP. For both the MLP layers, we set the hidden dimensionality to 16.

Predictor. We use a single layer MLP with no ReLU and BN as our predictor [Chen et al., 2020] to transform the graph embeddings of one augmented graph to another and vice-versa. We set the hidden dimensionality to 16 for the predictor layer.

Downstream classifier. We use a single fully-connected layer with a Sigmoid activation function in all our node-classification experiments. We set the hidden dimensionality of the fully-connected layer to 16.

Hyperparameters. For all experiments, we set the probability of perturbing a feature dimension to $p_n = 0.1$ and the probability with which an edge is dropped to $p_e = 0.001$. For training GNNs and their NIFTY-augmented counterparts (Sec. 6.1), we use an Adam optimizer with a learning rate of 1×10^{-3}, weight decay of 1×10^{-5}, and the number of epochs to 1000. For RobustGCN and FairGCN, all hyperparameters are set following the authors' guidelines.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

Kareem L Jordan and Tina L Freiburger. The effect of race/ethnicity on sentencing: Examining sentence type, jail length, and prison length. In *Journal of Ethnicity in Criminal Justice*. Taylor & Francis, 2015.

Jiachun Liao, Chong Huang, Peter Kairouz, and Lalitha Sankar. Learning generative adversarial representations (gap) under fairness and censoring constraints. *arXiv*, 2019.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *FAT*, 2019.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. In *ESA*, 2009.

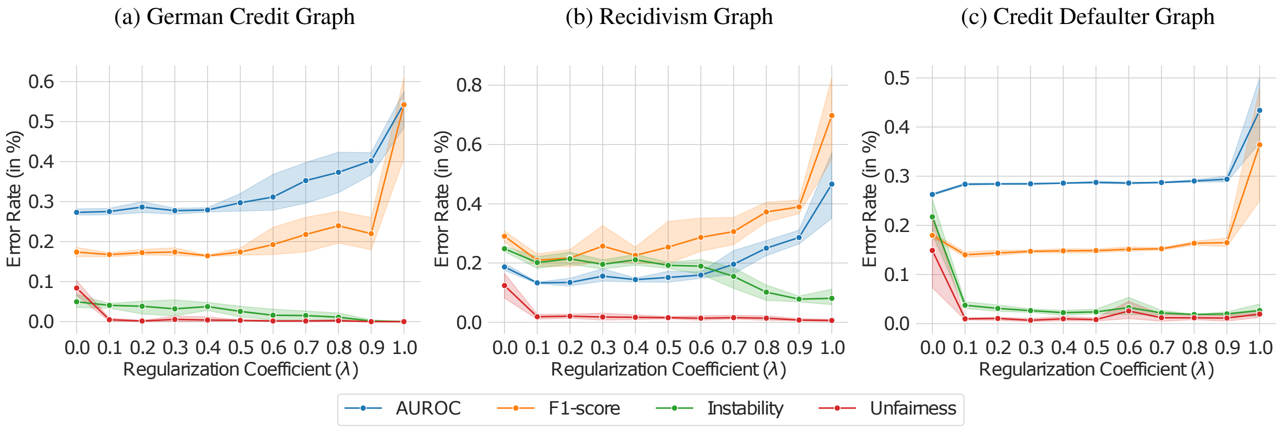

Figure 1: Effect of regularization coefficient on AUROC, F1-score, stability, and fairness in NIFTY-GIN on (a) the German credit graph, (b) the recidivism graph, and (c) the credit defaulter graph. With increasing the regularization coefficient on the self-supervised task the robustness and fairness score can reach 0% error.