



HAL
open science

Fair Graph Neural Network with Supervised Contrastive Regularization

Mahdi Tavassoli Kejani, Fadi Dornaika, Jean-Michel Loubes

► **To cite this version:**

Mahdi Tavassoli Kejani, Fadi Dornaika, Jean-Michel Loubes. Fair Graph Neural Network with Supervised Contrastive Regularization. 2024. hal-04537254

HAL Id: hal-04537254

<https://hal.science/hal-04537254>

Preprint submitted on 8 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Public Domain

Fair Graph Neural Network with Supervised Contrastive Regularization

Mahdi Tavassoli Kejani^{a,*}, Fadi Dornaika^{b,c} and Jean-Michel Loubes^a

^aInstitut de Mathématiques de Toulouse, Université Toulouse 3, Toulouse, France

^bUniversity of the Basque Country UPV/EHU, San Sebastian, Spain

^cIKERBASQUE, Basque Foundation for Science, Bilbao, Spain

ORCID (Mahdi Tavassoli Kejani): <https://orcid.org/0000-0002-7171-5746>, ORCID (Fadi Dornaika):

<https://orcid.org/0000-0001-6581-9680>, ORCID (Jean-Michel Loubes): <https://orcid.org/0000-0002-1252-2960>

Abstract

In recent years, Graph Neural Networks (GNNs) have made significant advancements, particularly in tasks such as node classification, link prediction, and graph representation. However, challenges arise from biases that can be hidden not only in the node attributes but also in the connections between entities. Therefore, ensuring fairness in graph neural network learning has become a critical problem. To address this issue, we propose a novel model for training fairness-aware GNN, which enhances the Counterfactual Augmented Fair Graph Neural Network Framework (CAF). Our approach integrates Supervised Contrastive Loss and Environmental Loss to enhance both accuracy and fairness. Experimental validation on three real datasets demonstrates the superiority of our proposed model over CAF and several other existing graph-based learning methods.

Keywords Graph Neural Networks (GNNs), fairness modeling, semi-supervised classification, counterfactual nodes, sensitive attributes

1 Modeling fairness using Graph Neural Networks

1.1 A review of fairness and Graph Neural Networks

Graph Neural Networks (GNNs) have emerged as a powerful paradigm for learning node representations on graphs, significantly advancing the state-of-the-art in tasks such as node classification, link prediction, and graph classification. By efficiently capturing the structural information within graphs, GNNs have found widespread application across various domains, including social network analysis, recommendation systems, drug discovery, and more.

Nowadays, Graph Neural Networks (GNNs) are a class of deep learning models designed to operate on data represented in graph form. Graphs consist of nodes (or vertices) and edges connecting these nodes, making them ideal for representing complex relational and structural information. The research of GNN began in [36] for a review. They faced several drawbacks [39, 20] such as model depth, scalability, convergence and more, but their construction have made significant progress over the years. However, challenges persist due to biases that may be embedded in node attributes or edges between nodes, resulting in unfair predictions with serious consequences. For example, in book recommendation with graph neural network, where the GNN methods could be biased towards suggesting books with male authors [8]. Therefore, it is crucial to ensure that these models and resulting representations are safe and reliable. Addressing bias and promoting fairness in GNNs are thus vital research areas.

However, as GNNs leverage historical data to learn representations, there is a growing concern that they may inadvertently perpetuate and even amplify existing biases present in the training data. This issue of fairness in GNNs is particularly critical because graphs often represent social structures, interactions, and personal relationships, where biases towards certain groups or individuals can have profound implications. For instance, in a social network, a biased model could lead to unfair recommendations, while in financial services, it might result in inequitable credit scoring.

Fairness has been studied over the last years with several methods to obtain fair decisions. Many methods try to balance the effect of unwanted correlations with variables that drive the bias in the sense that the decisions are different when the variable change while from moral, legal or ethical reasons, there should be no change. This variable is assumed to model a characteristic of the individuals that may convey bias leading to possible discrimination as defined in [10], [33], [4] or [5, 6] and references therein. In the literature on fairness in machine learning, such variable divides the individuals into a majority group corresponding to $S = 1$ and a minority group $S = 0$. The variable is often referred to as the sensitive attribute. The influence of this variable on the learnt algorithm may induce unwanted changes in the decision which can be measured using a statistical parity measure, or different level of performance for different subgroups measured by equality of odds measures.

* Corresponding Author. Email: mahdi.tavassoli-kejani@univ-toulouse.fr.

Many authors propose several ways to tackle this issue. We refer for instance to [16], [42], [35], [17] and references therein, which propose several techniques to add fairness constraint when learning the algorithm. Another general way to achieve fairness is to use a latent space which disentangles the effect of the characteristics and the sensitive variables. This idea is developed for instance in [2], [43]. Such ideas are very similar to the notions of counterfactual in fairness as presented in [29], [14] where the authors propose to model how changes in the sensitive variable can be modeled as a causal intervention enabling to respond to the question *what-if* the sensitive variable had been different and what would have been the corresponding decision. This counterfactual definition of fairness implies the so-called statistical parity fairness and will be related to the notion of fairness for graph neural networks.

Recent research has begun to address these concerns by investigating the sources of bias in graph data and developing methodologies to mitigate discriminatory biases in GNN-based applications. A notable direction in this field includes the identification and quantification of bias in node embeddings, exploring how structural properties of graphs, such as homophily and community structure, can contribute to bias amplification. Other publications offer insights on how to address fairness in the construction of the GNNs. For instance, [38, 23, 24] introduced the concept of fairness in graph representation learning, highlighting the need for fairness-aware algorithms in GNNs. Furthermore, research by [25] proposed a framework for fair node classification, which adapts adversarial debiasing techniques to the graph setting, demonstrating the potential for achieving fairness in GNN predictions. Optimal transport methods popular in fairness as in [17] have been successfully applied for graph and edges prediction tasks in [30].

Efforts to mitigate bias in graph-structured data have led to the development of fairness-aware GNNs [13, 19, 39]. For instance, FairGNN [12] employs adversarial learning to ensure node representations remain independent of sensitive attributes. Another notable approach is EDITS [15], which concentrates on addressing bias in both the adjacency matrix and data matrix as part of the pre-processing stage. Additionally, Fairwalk [34] utilizes a fair random walk method to learn fair network embeddings.

Recent studies have turned their attention to attaining counterfactual fairness within graph data. In this context, a classification prediction for an individual is considered counterfactually fair when it stays consistent across both the actual and counterfactual worlds. For instance, NIFTY [1] maximizing the similarity between representations of the original nodes in the graph and their counterparts in the augmented graph. These counterparts are generated by either slightly perturbing the original node attributes and edges or considering counterfactuals of the original nodes where the value of the sensitive attribute is modified and GEAR [31] adopts GraphVAE [28] to generate counterfactuals corresponding to perturbations on each node’s and their neighbors’ sensitive attributes. It then enforces fairness by minimizing the discrepancy between the representations learned from the original graph and the counterfactuals for each node.

These approaches [1, 28] can yield counterfactuals that are potentially unrealistic, which, in turn, have the capacity to disrupt the underlying semantic structure and decrease the performance of the model. However, finding realistic counterfactuals presents several challenges: (i) Graph data is quite complex, making it infeasible to directly identify counterfactuals in the original data space. Additionally, some guidance or rules are needed to facilitate the discovery of counterfactuals. (ii) Achieving graph counterfactual fairness requires learned representations to be invariant to sensitive attributes and information causally influenced by sensitive attributes.

In CAF [18], the authors address the aforementioned challenges by constructing a Structural Causal Model to identify counterfactuals and learn disentangled representations. They divided these representations into content and environmental information parts, ensuring that the content information remains informative for labels and invariant to sensitive attributes, while the environmental information depends on sensitive attributes. Despite constructing a fair model based on these principles using realistic counterfactuals, experimental results from CAF indicate that while fairness is improved, performance is decreased.

The results presented in the CAF paper indicate that while improvements were made in fairness, the price for achieving fair decisions leads to a drastic decrease of the accuracy of the decisions. This motivated us to extend this model by introducing two extra constraints aimed at simultaneously improving fairness and accuracy, leading to a better balance preserving the efficiency of the algorithm. Our main contributions are:

1. We use the Supervised Contrastive Loss to encourage similar content information with similar labels to have similar representations, irrespective of sensitive attributes.
2. We use the Environmental Loss to encourage dissimilarity in environment information when the sensitive attributes are dissimilar.
3. We propose a novel framework by extending the CAF framework and utilizing the Supervised Contrastive Loss and Environmental Loss. We demonstrate the effectiveness of our model through experiments on real-world datasets and comparison with other methods.

1.2 Notations

In this section, we present the notations employed in this paper. We assume we observe $i = 1, \dots, n$ individuals with characterizes $x_i \in \mathbb{R}^d$ for $d > 0$. Let define the matrix of observations \mathcal{X} formed by combining vectors $[x_1, x_2, \dots, x_n]$ in $\mathbb{R}^{n \times d}$. The aim of our study is to forecast a binary label $Y \in \{0, 1\}$ which is only observed for some individuals. In addition, the individuals are also characterized by an observed binary variable $S \in \{0, 1\}$. The influence of the sensitive variable on the predictions is at the heart of our study. Our aim is to mitigate its effect and build a latent space where the effects of the features and the sensitive variables are separated, enabling bias control and bias mitigation.

We will model the relationships between individuals using a graph where the nodes are the observed individuals. A graph is defined as a couple $G = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} composed by a set of nodes of size $|\mathcal{V}| = n$, \mathcal{E} and a set of edges. The nodes are connected by edges and the connections are modeled using a matrix $\mathcal{A} \in \mathbb{R}^{n \times n}$ known as the adjacency matrix of the graph G . As usual $\mathcal{A}_{i,j} = 1$ if edge $i \rightarrow j$ exists, otherwise $\mathcal{A}_{i,j} = 0$. Further we assume that G is undirected and unweighted. We refer for instance to [7] for definitions of graphs. The individuals $i = 1, \dots, n$ are characterized by their features denoted $x_i \in \mathbb{R}^d$ that can be used or not in the construction of the initial graph.

We are in a semi-supervised setting where $\{1, \dots, n\} = \mathcal{L} \cup \mathcal{U}$, with $\mathcal{L} = \{1, \dots, l\}$ contains the indices of the labeled samples while

$\mathcal{U} = \{l + 1, \dots, l + u\}$ is the set of indices of the unlabeled samples.

In this paper we only present the case of a binary sensitive attribute. Yet, following the works [31] and [18] our model seamlessly extends to handle various types of weighted graphs and sensitive attributes, not only the binary case.

2 Contrastive Loss for Fair Graph Neural Networks

2.1 Setting a Contrastive Graph Neural Network

Traditional methods for learning node representations utilize an encoder $f_\theta(\cdot) : \mathcal{X}^{n \times d} \times \mathcal{A}^{n \times n} \rightarrow \mathcal{H}^{n \times d'}$ to map each node to a latent lower dimensional representation so that it preserves both node attributes and local graph structure information. Here, f_θ represents graph neural network (GNN) architectures, such as GCN [27] and GraphSAGE [21], and \mathcal{H} denotes the latent representations.

In the study by Guo et al. [18], their objective is to ensure that label predictions are independent from chosen sensitive attributes, hence obtaining so-called *fair* prediction. For this, they introduce the Counterfactual Augmented Fair Graph Neural Network Framework (CAF) for learning fair node representations. Fairness is achieved by imposing several constraints and mainly partitioning the latent representation \mathcal{H} into two distinct components: C (content information) and E (environmental information). This partition of \mathcal{H} is formally represented as follows:

$$f_\theta(\mathcal{X}, \mathcal{A}) = \mathcal{H} = [C, E] \quad (1)$$

The first d_c columns of H constitute the content information matrix C , which contains information relevant for label prediction while remaining unaffected by changes of sensitive attributes. This content information (C) is the only one that will be used for prediction. Actually a classifier f_ϕ parameterized by ϕ is learnt by using the content information as input while forecasting the class distribution of each node as follows:

$$f_\phi(c) = \hat{y} \quad (2)$$

While the next d_e columns form the environment information matrix E which is designed to capture environmental information primarily dependent on sensitive attributes. Notably, d_c is equal to d_e .

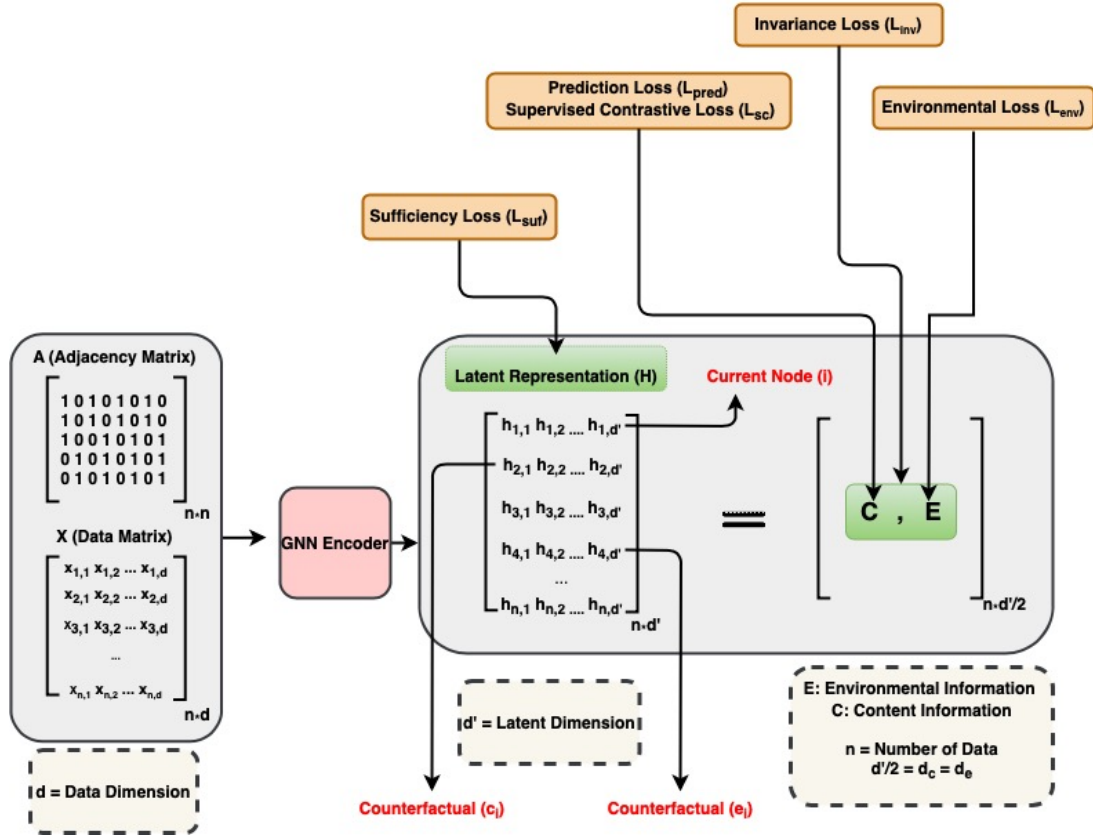


Figure 1: An illustration of our proposed framework

For a given nodes i we can consider two notions of counterfactual nodes: we can either look at the closest nodes with similar features and the same outputs of the algorithms but with distinct sensitive attributes or the closest nodes with similar features, the same sensitive attributes

but different outputs of the algorithm. We want to design the embedding in a such a way that, in the first case, the changes in the features h will mainly be changes with respect to environment part e since this part should be related to the sensitive attribute. On the contrary, we expect, in the second case, that the changes will be driven by the c part of the features, i.e the content which explain the decision. Hence we will use the following notations

$$h_i^e \in \operatorname{argmin}_{h_j \in H} \{\|h_i - h_j\|_2^2 \mid \hat{y}_i = \hat{y}_j, s_i \neq s_j\} \quad (3)$$

$$h_i^c \in \operatorname{argmin}_{h_j \in H} \{\|h_i - h_j\|_2^2 \mid \hat{y}_i \neq \hat{y}_j, s_i = s_j\} \quad (4)$$

and we note $h_i^e = (c_i^e, e_i^e)$ and respectively $h_i^c = (c_i^c, e_i^c)$.

We point out that the corresponding counterfactuals are not uniquely defined for each node. We denote by $\mathcal{K}^e(i)$ and $\mathcal{K}^c(i)$ the sets of counterfactuals of a node i and we set K the same number of chosen counterfactuals.

We will build a Contrastive Counterfactually Fair Graph Neural Network by considering the following losses. The first term $\mathcal{L}_{\text{pred}}$ is the usual Cross Entropy Loss, referred to as the prediction loss.

$$\mathcal{L}_{\text{pred}} = \frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} \ell(\hat{y}_i, y_i). \quad (5)$$

Note also that computing the counterfactuals requires knowing the estimation of the labels for all points. Hence, due to the semi-supervised setting we have a limited number of labels in the training set for counterfactual guidance, and thus the model is pre-trained to obtain pseudo-labels \hat{y} by applying the following loss function to Equation (2).

To learn fair latent representations the loss is designed to achieve the decomposition into content and environment variables. Namely the purpose is that when the sensitive attribute s_i of a node i is flipped to $1 - s_i$, its content representation c_i is expected to remain constant, while its environment representation e_i changes correspondingly, resulting in the formation of the counterfactual h_i^e . Similarly, when the label y_i of a node i is flipped while keeping the sensitive attribute s_i unchanged, the content representation c_i is changed, while the environment representation e_i remain constant, forming h_i^c .

The second term, \mathcal{L}_{inv} , is the invariance loss, defined as follows:

$$\mathcal{L}_{\text{inv}} = \frac{1}{|\mathcal{V}| \cdot K} \sum_{v_i \in \mathcal{V}} \sum_{k=1}^K [\operatorname{dis}(c_i, c_i^{e_k}) + \operatorname{dis}(e_i, e_i^{c_k}) + \gamma K \cdot |\cos(c_i, e_i)|]$$

where $\operatorname{dis}(\cdot, \cdot)$ is a distance metric, such as the cosine distance and $L2$ distance. $|\cos(\cdot, \cdot)|$ represents the absolute value of the cosine similarity, which imposes an orthogonality constraint, and γ is the hyper-parameter that controls it. This loss promotes fairness of the embedding.

The third term, \mathcal{L}_{suf} , the sufficiency loss, which is defined in order to preserve the graph information and ensure that the latent representation has the same topological property as the initial input graph. It is defined as follows:

$$\mathcal{L}_{\text{suf}} = \frac{1}{|\mathcal{A}| + |\mathcal{A}^-|} \times \sum_{(i,j) \in \mathcal{A} \cup \mathcal{A}^-} -a_{i,j} \log p_{i,j} - (1 - a_{i,j}) \log(1 - p_{i,j})$$

where \mathcal{A}^- is the set of sampled negative edges. If nodes i and j are connected, $a_{i,j} = 1$; otherwise, $a_{i,j} = 0$. Additionally, $p_{i,j} = \sigma(h_i h_j^T)$ represents the link existence probability, where σ denotes the Sigmoid function.

Note that Guo et al. [18], uses the following loss function to learn fair classification

$$\min_{\theta, \phi} \mathcal{L} = \mathcal{L}_{\text{pred}} + \alpha \mathcal{L}_{\text{inv}} + \beta \mathcal{L}_{\text{suf}} \quad (6)$$

where θ and ϕ are GNN encoder and prediction head parameters, and α and β are hyper-parameters controlling invariance and sufficiency. Although CAF can enhance fairness, it tends to reduce accuracy since it ocuses solely on splitting content and environment information, ensures that the sensitive attribute does not impact node classification tasks.

2.2 Improvements using Contrastive Loss functions

In our study we go on improving their model by incorporating two losses that will be referred to as the Supervised Contrastive Loss and the Environmental Loss and prove that they enable to mitigate fairness without harming accuracy. Our architecture is displayed in Figure (1).

Supervised Contrastive Loss, \mathcal{L}_{sc} , is a crucial element in Supervised Contrastive Learning [26]. It works by encouraging similar data points to have similar representation while pushing dissimilar data point apart in the Latent space. Using this property, this loss has been successfully used to model counterfactuals since it encourages separation of the points in the embedding space based on the choice of a relevant similarity. For instance in [37], it enables to represent time series sharing into an embedding space made of two components, one providing a common

pattern while the other models the difference between the functions. In the context of graph representation, we apply this loss specifically to the content information (C) of a latent representation, defines as follows

$$\mathcal{L}_{sc} = \sum_{i=1}^n \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp\left(\frac{c_i \cdot c_p}{\tau}\right)}{\sum_{a \in \mathcal{V}(i)} \exp\left(\frac{c_i \cdot c_a}{\tau}\right)} \quad (7)$$

where $\mathcal{P}(i) \equiv \{p \in \mathcal{V}(i) : \hat{y}_p \equiv \hat{y}_i\}$ is the set of indices of all positive samples (those with the class labels). Nodes sharing similar content and similar labels tend to have similar embedding in the latent space.

In our proposed model, we apply this loss during the pretrained phase using labeled data. However, during the training phase, we apply this loss to all training data, utilizing pseudo-labels generated during the pretrained process.

The Environmental Loss, \mathcal{L}_{env} , works by encouraging node has dissimilar sensitive attribute to have dissimilar environment information in the latent space. \mathcal{L}_{env} is defined as follows:

$$\mathcal{L}_{env} = \frac{1}{n} \sum_{i=1}^n \frac{1}{K'} \sum_{j=1}^{K'} \text{dis}(e_i, e_j) \quad (8)$$

Where K' denotes the number of nodes that have dissimilar sensitive attributes but whose latent representations are close to the current node.

In our proposed model, we apply this loss during the pretrained and training phase.

According to the above definition, the loss function of our proposed model for the pretraining phase is defined as follows:

$$\min_{\theta, \phi} \mathcal{L} = \mathcal{L}_{pred} + \omega \mathcal{L}_{sc} - \eta \mathcal{L}_{env} \quad (9)$$

and for the training phase is :

$$\min_{\theta, \phi} \mathcal{L} = \mathcal{L}_{pred} + \alpha \mathcal{L}_{inv} + \beta \mathcal{L}_{suf} + \omega \mathcal{L}_{sc} - \eta \mathcal{L}_{env} \quad (10)$$

All the parameters $\omega, \eta, \alpha, \beta, \gamma$ are trade-off parameters that balance the influence of all terms. Their influence and the choice is described in the following section.

As mentioned, we divide the latent representation into content and environmental information parts. Fairness and performance improve when the content information contains no sensitive attribute-related information, and all such information exists in the environmental part. Therefore, we utilize the Supervised Contrastive Loss and Environmental Loss to enforce that each part has related information.

3 Experiments

Table 1: The evaluation of node classification and group fairness performance across real-world datasets

Dataset \ Method	Metric	GCN	GraphSAGE	GIN	FairGNN	EDITS	GEAR	CAF	SCCAF
GERMAN	AUC (\uparrow)	57.38 \pm 2.85	62.14 \pm 5.35	59.16 \pm 5.98	54.28 \pm 3.99	50.75 \pm 4.06	44.35 \pm 5.88	59.76 \pm 5.88	66.07 \pm 2.97
	F1 (\uparrow)	65.51 \pm 1.45	73.15 \pm 14.18	72.40 \pm 3.91	62.40 \pm 9.14	62.34 \pm 6.09	50.78 \pm 21.61	81.38 \pm 1.22	82.38 \pm 0.11
	Δ_{SP} (\downarrow)	5.13 \pm 4.30	4.09 \pm 4.56	6.82 \pm 1.98	8.76 \pm 7.82	7.99 \pm 5.88	8.38 \pm 4.93	3.93 \pm 3.05	2.4 \pm 3.11
	Δ_{EO} (\downarrow)	5.71 \pm 5.85	4.62 \pm 4.57	7.03 \pm 3.94	10.44 \pm 8.73	6.92 \pm 6.02	8.29 \pm 4.37	3.73 \pm 1.05	1.1 \pm 1.34
BAIL	AUC (\uparrow)	93.69 \pm 1.27	97.17 \pm 0.87	94.77 \pm 0.82	94.87 \pm 1.07	92.36 \pm 1.40	86.86 \pm 1.65	97.45 \pm 0.09	98.01 \pm 0.49
	F1 (\uparrow)	84.57 \pm 1.34	93.40 \pm 1.60	89.06 \pm 1.67	87.57 \pm 2.67	82.75 \pm 2.43	79.00 \pm 2.47	94.43 \pm 0.63	95.48 \pm 1.88
	Δ_{SP} (\downarrow)	6.88 \pm 1.56	7.08 \pm 1.17	7.14 \pm 1.59	7.40 \pm 0.76	7.31 \pm 1.15	7.95 \pm 1.55	6.45 \pm 1.48	7.54 \pm 1.09
	Δ_{EO} (\downarrow)	1.48 \pm 0.93	1.14 \pm 1.19	1.54 \pm 1.26	1.12 \pm 0.76	1.94 \pm 1.35	3.16 \pm 1.58	0.87 \pm 1.21	0.64 \pm 0.84
CREDIT	AUC (\uparrow)	67.36 \pm 2.31	71.92 \pm 1.95	67.956 \pm 6.43	63.19 \pm 3.67	66.64 \pm 0.35	69.17 \pm 0.52	65.75 \pm 2.38	70.03 \pm 1.39
	F1 (\uparrow)	80.06 \pm 2.28	78.64 \pm 0.71	72.28 \pm 6.65	74.62 \pm 1.07	84.42 \pm 3.02	78.84 \pm 1.02	86.36 \pm 0.26	86.83 \pm 0.15
	Δ_{SP} (\downarrow)	12.93 \pm 5.36	10.54 \pm 4.42	8.79 \pm 6.86	9.39 \pm 2.65	3.16 \pm 1.61	14.298 \pm 1.41	3.43 \pm 1.67	3.47 \pm 2.22
	Δ_{EO} (\downarrow)	11.40 \pm 5.92	9.46 \pm 4.36	8.87 \pm 7.50	8.85 \pm 3.12	1.97 \pm 1.43	13.24 \pm 2.65	1.67 \pm 1.93	1.52 \pm 1.84

In this section, we assess the proposed model’s performance on diverse tabular real-world datasets, encompassing German credit [3], Bail [22] and Credit Defaulter [41].

3.1 Datasets

We used the following datasets:

1. **German Credit [3]:** The dataset consists of 1000 individual clients from a German bank, and these individuals, acting as nodes, are linked when there is a significant similarity in their credit accounts. The main objective is to evaluate the credit risk level for each individual, emphasizing the sensitive attribute of ‘gender’.
2. **Bail [22]:** The dataset comprises 18,876 individuals granted bail by US state courts between 1990 and 2009. These individuals, serving as nodes, are linked based similarities in their demographics and prior criminal histories. The primary goal is to categorize defendants as either on bail or not, with a focus on the sensitive attribute ‘race’.
3. **Credit Defaulter [41]:** The dataset includes 30,000 individual credit card holders, where these individuals, acting as nodes, are linked based on high similarity in payment information. The main objective is to predict whether an individual will default on credit card payments, with particular attention to the sensitive attribute of ‘age’.

3.2 Experimental setup

We compare the proposed SCCAF framework with state-of-the-art plain node classification methods, including GCN [27], GraphSAGE [21], and GIN [40]. Additionally, we assess its performance against fair node classification methods (FairGNN [12], EDITS [15]), as well as graph counterfactual fairness method (GEAR [32]) and fair node representations for graph counterfactual fairness method (CAF [18])

For all databases, we adhere to the train/validation/test split methodology described in [18]. However, to ensure a more robust assessment of our method alongside others, we compute the average results obtained from five random splits of the data into train/validation/test sets.

Each method under comparison is associated with its own set of hyperparameters, which undergo optimization through a grid search. Subsequently, the model demonstrating the best performance for each method is selected. This standardized approach is consistently applied across all methods being compared. Additionally, the learning rate is fixed at 0.01, and the random seed is kept constant.

For our proposed method, the parameters are set as follows: K is selected from {4, 6, 10}, K' is selected from {10, 20}, α is selected from {0.2, 0.1, 0.07}, β is selected from {0.1, 0.05, 0.01}, γ is set to 0.01, ω is set to 0.01, and η is selected from {0.1, 0.03}

Moreover, to ensure impartial evaluations, all methods undergo testing using identical train/validation/test splits. This uniform methodology ensures fair and equitable comparisons across all evaluated methodologies.

3.3 Related fairness metrics

Researchers assess fairness for Graph Neural Networks across various fundamental levels. In their work [9], the authors classify fairness evaluation metrics for Graph Neural Networks into prediction-level metrics, graph-level metrics, neighborhood-level metrics and embedding-level metrics. In our work, we focus on prediction-level metrics, which we explain in this section.

Statistical Parity (SP) [12]: The statistical parity (SP) metric, also known as demographic parity (DP), requires that predictions made by the model are independent of the sensitive attribute s . It is defined as follows:

$$\Delta_{SP} = |\mathcal{P}(\hat{y}_u = 1 | s = 0) - \mathcal{P}(\hat{y}_u = 1 | s = 1)| \quad (11)$$

Where $\Delta_{SP} = 0$ implies that all groups have the same selection rates, indicating complete fairness. Statistical parity measures the gap in preferential treatment between different groups. However, it's important to note that Δ_{SP} does not consider whether individuals are qualified or not, as it does not take into account the ground-truth label Y .

Equal Opportunity (EQ) [12]: This principle dictates that the probability of a positive outcome being predicted for instances in a positive class should be the same for all subgroup members. It is defined as follows:

$$\Delta_{EO} = |\mathcal{P}(\hat{y} = 1 | y = 1, s = 0) - \mathcal{P}(\hat{y} = 1 | y = 1, s = 1)|$$

Where $\Delta_{EQ} = 0$ implies equal true positive rates across subgroups, indicating fairness.

3.4 Method comparison

We assess the model's performance. Table (1) presents the AUC and F1 score as percentages, along with their standard deviations across five random splits, providing insights into node classification proficiency. In terms of fairness evaluation, guided by [11], we employ two widely recognized group fairness metrics specifically, statistical parity (SP) Δ_{SP} and equality parity (EO) Δ_{EO} .

3.5 Effect of the proposed SCCAF for fair classification task

The results in Table (1) demonstrate that our proposed method can improve performance and fairness simultaneously. In the German database, we enhance the AUC and F1 metrics, related to performance, from 59.76 and 81.38 in the CAF method to 66.07 and 82.38, respectively, while Δ_{SP} and Δ_{EQ} , related to fairness metrics, decrease from 3.93 and 3.73 to 2.4 and 1.1. In the Bail database, we enhance the AUC and F1 metrics, related to performance, from 97.45 and 94.43 in the CAF method to 98.01 and 95.48, respectively, and Δ_{EQ} metric decreases from 0.87 to 0.64. In the Credit database, we enhance the AUC and F1 metrics, related to performance, from 65.75 and 86.36 in the CAF method to 70.03 and 86.83, respectively, and Δ_{EQ} metric decreases from 1.67 to 1.52. We can observe a significant improvement over other methods. Therefore, we can conclude that if we can separate the content and environmental information in the best ways, both performance and fairness increase simultaneously.

3.6 Effect of the Supervised Contrastive Loss

The proposed method incorporates two parameters: ω , which determines the weight of the Supervised Contrastive Loss in the total loss, and η , which determines the weight of the Environmental Loss in the total loss. Figure (2) illustrates the accuracy and fairness performance as a function of ω while other parameters are fixed (with $\alpha = 0.2$, $\beta = 0.1$, $\gamma = 0.01$, and $\eta = 0.1$). Similarly, it shows the accuracy and fairness performance as a function of η while other parameters are fixed (with $\alpha = 0.2$, $\beta = 0.1$, $\gamma = 0.01$, and $\eta = 0.1$), using the test data from the Bail database. Increasing ω and η generally leads to a trade-off between prediction performance and fairness, although the changes are not as significant as those observed in CAF. Therefore, this paper aims to achieve our goal of improving accuracy and fairness simultaneously.

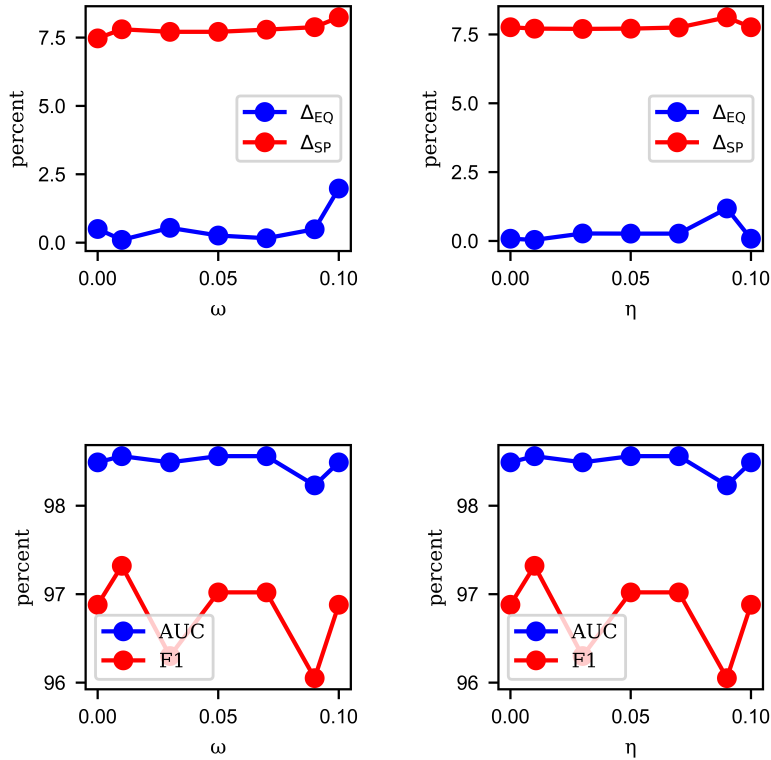


Figure 2: Δ_{EQ} , Δ_{SP} , AUC, and F1 percentages of the proposed model SCCAF as functions of the parameter ω and η in the Bail databases

4 Conclusion

In this paper, we introduced a novel framework called SCCAF, which enhances the CAF model for learning fair node representations. Similarly to CAF model, the main objective is still to achieve fairness by imposing a notion of distributional independence in the latent space. This is usually achieved by partitioning the latent representation into two parts : a content and an environmental information part. The content information contains no sensitive attribute-related data, while all such information resides in the environmental part. The crucial aspect lies in how we define the loss function to ensure both separation of the information and yet effective learning of both the content and environmental information parts with relevant data. The loss function defined in the CAF method is not sufficient leading to separation but poor accuracy due to a too large loss of information.

Hence we proposed to first use a Supervised Contrastive Loss, which yet requires to estimate the pseudo-labels of all samples. This approach improves the learning of the content information part. Furthermore, to improve the relevance of the environmental latent space and its connection with the sensitive variable, we incorporated an Environmental Loss to ensure that for different sensitive attributes, their environmental part also differs. By adding these losses to the CAF loss, we achieved our goal effectively, leading to increased fairness and accuracy simultaneously. The results table and figures demonstrate that our proposed method outperforms CAF and other methods.

References

- [1] C. Agarwal, H. Lakkaraju, and M. Zitnik. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in Artificial Intelligence*, pages 2114–2124. PMLR, 2021.
- [2] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019.
- [3] A. Asuncion and D. Newman. Uci machine learning repository, 2007.
- [4] S. Barocas, M. Hardt, and A. Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2, 2017.
- [5] P. Besse, C. Castets-Renard, A. Garivier, and J.-M. Loubes. Can everyday ai be ethical? machine learning algorithm fairness. *Machine Learning Algorithm Fairness (May 20, 2018)*. *Statistiques et Société*, 6(3), 2019.
- [6] P. Besse, E. del Barrio, P. Gordaliza, J.-M. Loubes, and L. Rissler. A survey of bias in machine learning through the prism of statistical parity. *The American Statistician*, pages 1–11, 2021.
- [7] J. A. Bondy and U. S. R. Murty. *Graph theory*. Springer Publishing Company, Incorporated, 2008.
- [8] M. Buyl and T. De Bie. Debayes: a bayesian method for debiasing network embeddings. In *International Conference on Machine Learning*, pages 1220–1229. PMLR, 2020.
- [9] A. Chen, R. A. Rossi, N. Park, P. Trivedi, Y. Wang, T. Yu, S. Kim, F. Deroncourt, and N. K. Ahmed. Fairness-aware graph neural networks: A survey. *ACM Transactions on Knowledge Discovery from Data*, 2023.
- [10] A. Chouldechova and A. Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

- [11] E. Dai and S. Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM '21*, page 680–688, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382977. doi: 10.1145/3437963.3441752. URL <https://doi.org/10.1145/3437963.3441752>.
- [12] E. Dai and S. Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 680–688, 2021.
- [13] E. Dai, T. Zhao, H. Zhu, J. Xu, Z. Guo, H. Liu, J. Tang, and S. Wang. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *arXiv preprint arXiv:2204.08570*, 2022.
- [14] L. De Lara, A. González-Sanz, N. Asher, and J.-M. Loubes. Transport-based counterfactual models. *arXiv preprint arXiv:2108.13025*, 2021.
- [15] Y. Dong, N. Liu, B. Jalaian, and J. Li. Edits: Modeling and mitigating data bias for graph neural networks. In *Proceedings of the ACM Web Conference 2022*, pages 1259–1269, 2022.
- [16] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.
- [17] P. Gordaliza, E. Del Barrio, G. Fabrice, and J.-M. Loubes. Obtaining fairness using optimal transport theory. In *International conference on machine learning*, pages 2357–2365. PMLR, 2019.
- [18] Z. Guo, J. Li, T. Xiao, Y. Ma, and S. Wang. Towards fair graph neural networks via graph counterfactual. Oct. 2023. doi: 10.1145/3583780.3615092. URL <http://dx.doi.org/10.1145/3583780.3615092>.
- [19] Z. Guo, T. Xiao, C. Aggarwal, H. Liu, and S. Wang. Counterfactual learning on graphs: A survey. *arXiv preprint arXiv:2304.01391*, 2023.
- [20] A. Gupta, P. Matta, and B. Pant. Graph neural network: Current state of art, challenges and applications. *Materials Today: Proceedings*, 46:10927–10932, 2021.
- [21] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.
- [22] K. L. Jordan and T. L. Freiburger. The effect of race/ethnicity on sentencing: Examining sentence type, jail length, and prison length. *Journal of Ethnicity in Criminal Justice*, 13(3):179–196, 2015.
- [23] J. Kang and H. Tong. Fair graph mining. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4849–4852, 2021.
- [24] J. Kang and H. Tong. Algorithmic fairness on graphs: Methods and trends. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4798–4799, 2022.
- [25] A. Khajehnejad, M. Khajehnejad, M. Babaei, K. P. Gummadi, A. Weller, and B. Mirzasoleiman. Crosswalk: Fairness-enhanced node representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11963–11970, 2022.
- [26] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.
- [27] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [28] T. N. Kipf and M. Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.
- [29] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076. Curran Associates, Inc., 2017.
- [30] C. Laclau, I. Redko, M. Choudhary, and C. Llargeron. All of the fairness for edge prediction with optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1774–1782. PMLR, 2021.
- [31] J. Ma, R. Guo, M. Wan, L. Yang, A. Zhang, and J. Li. Learning fair node representations with graph counterfactual fairness. Feb. 2022. doi: 10.1145/3488560.3498391. URL <http://dx.doi.org/10.1145/3488560.3498391>.
- [32] J. Ma, R. Guo, M. Wan, L. Yang, A. Zhang, and J. Li. Learning fair node representations with graph counterfactual fairness. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 695–703, 2022.
- [33] L. Oneto, M. Donini, G. Luise, C. Ciliberto, A. Maurer, and M. Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20*, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.
- [34] T. Rahman, B. Surma, M. Backes, and Y. Zhang. Fairwalk: Towards fair graph embedding. 2019.
- [35] L. Risser, A. G. Sanz, Q. Vincenot, and J.-M. Loubes. Tackling algorithmic bias in neural-network classifiers using wasserstein-2 regularization. *J. Math. Imaging Vis.*, 64(6):672–689, 2022.
- [36] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80, 2008.
- [37] W. Todo, M. Selmani, B. Laurent, and J.-M. Loubes. Counterfactual explanation for multivariate times series using a contrastive variational autoencoder. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [38] S. Venkatasubramanian, C. Scheidegger, S. Friedler, and A. Cluset. Fairness in networks: social capital, information access, and interventions. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4078–4079, 2021.
- [39] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.
- [40] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks?, 2019.
- [41] I.-C. Yeh and C.-h. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480, 2009.
- [42] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *International Conference on Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.
- [43] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333, 2013.