Appendix for the paper: The KL-Divergence between a Graph Model and its Fair I-Projection as a Fairness Regularizer

1 Hyperparameters and Implementation

As stated in the paper, hyperparameter tuning was minimal. In this section, we nevertheless provide, for each method, additional details on the choice of hyperparameters and implementations. The source code is also provided in the supplementary submission.

MaxEnt

The MaxEnt model is used in both the CNE and DeBayes models as a prior distribution. We therefore used the code available for CNE (available here) as a starting point for developing a *PyTorch* version of this model. The optimisation was done with *LBFGS*, with at most 100 iterations.

Dot-Product

Due to the simplicity of the Dot-Product model, it was implemented from scratch in *PyTorch*. A dimensionality of 128 performed the best among the considered values $\{8, 16, 128\}$. The model was optimised with *Adam* and at a learning rate of 0.01 with 100 iterations.

CNE

The CNE implementation was based on the code from DeBayes, since the latter is a fairness adaptation of CNE. Our *PyTorch* modification was made similar to Dot-Product, except that it uses the distance between node embeddings instead of the actual dot product, and it uses MaxEnt as a prior distribution. As hyperparameters, we used the default values as listed for DeBayes: dimensionality of 8, learning rate of 0.1 and s_2 value of 16. However, we found the method already converged with 200 epochs, as opposed to the default 1000.

GAE

For the GAE implementation, we closely followed the source code here. We used similar parameters as Dot-Product, but found a dimensionality of 16 achieved better validation scores.

CFC

We applied the CFC method to link prediction by using it as a dot-product decoder. As such, the hyperparameters are the same as for Dot-Product, though we stayed as close as possible to the original implementation available here.

DeBayes

Our implementation of DeBayes is based on the publicly available version here. However, we used our own *PyTorch* version of CNE in order to compare both methods easily. The DeBayes method has no other hyperparameters besides those listed for CNE.

2 Additional Measures

In addition to the DP and EO measures, we evaluated the considered methods on two additional fairness measures.

Rank DP (RDP)

As stated in the paper, we computed the DP measure as the maximum difference between the mean link prediction probability of each subgroup combination, proposed in prior work [2]. A possible drawback of this approach is that an algorithm can reduce its DP score by simply making all prediction probabilities very similar. For example, a dot-product decoder may give privileged vertex pairs a link probability of 0.9, and unprivileged pairs a score of 0.5. Such a method would have a DP score of 0.4. However, e.g. by rescaling the embeddings, a similar dot-product decoder may predict scores of 0.6 and 0.5 for privileged and unprivileged pairs respectively, which would result in a DP score of 0.1. Both methods provide total separation in the *ranks* of the scores between subgroups, so it could be argued that they are both extremely (yet equally) discriminatory.

Taking inspiration from prior work [3], consider the AUC score, computed over the link prediction scores on the test set, when pairs of vertices of sensitive groups V_s and V_t respectively are the positive class and vertices with a different sensitive attribute combination are the negative class. Let the *Rank Demographic Parity* (RDP) score be the maximum of those AUC scores, over all sensitive value pairs $(s, t) \in S$. Clearly, when the prediction scores are completely separated between groups as in the earlier example, the RDP score would equal 1.

In the results reported in Sec. 3, the RDP score heavily correlated with the DP score, suggesting that the latter is already an adequate measure of demographic parity in our experiment setup.

Representation Bias (RB)

The baseline methods DeBayes [2] and CFC [1] focus on making the embeddings themselves less biased. While this was not the direct aim of our fairness regularizer, we evaluate all considered methods by also measuring the *Representation Bias* (RB) [2], i.e. the maximum AUC score that a logistic regression classifier achieves when it attempts to predict the sensitive attribute value of a vertex' embedding.

We report the scores using such an evaluation measure in Sec. 3 and validate that DeBayes and CFC indeed succeed in debiasing their embeddings, as reported in their respective papers. The hyperparameters of the classifier are tuned on a validation set of vertices among all vertices in the graph. In some cases, e.g. for Dot-Product (DP), the optimal hyperparameters are those that cause extreme regularisation, causing the RB scores to be 0.5.

3 Results in Table Format

The full experimental results, which were displayed as AUC-fairness trade-offs in the paper, are again displayed in Tab. 1, 2 and 3. In addition to the DP and EO fairness measures, the RDP and RB scores as described in Sec. 2 are also listed.

Table 1. Mean ± standard deviation of results on the POLBLOGS dataset.

METHOD	AUC	DP	EO	RDP	RB
CFC	0.938 ± 0.0047	0.145 ± 0.0151	0.204 ± 0.0211	0.664 ± 0.0154	0.977 ± 0.0049
CFC ($\lambda = 10$)	0.920 ± 0.0407	0.106 ± 0.0346	0.228 ± 0.0766	0.668 ± 0.0336	0.940 ± 0.1194
CFC ($\lambda = 100$)	0.789 ± 0.0578	0.058 ± 0.0463	0.096 ± 0.0662	0.626 ± 0.0442	0.859 ± 0.1188
CFC ($\lambda = 1000$)	0.544 ± 0.0297	0.010 ± 0.0071	0.015 ± 0.0101	0.520 ± 0.0135	0.595 ± 0.0482
CNE	0.962 ± 0.0022	0.077 ± 0.0033	0.226 ± 0.0079	0.679 ± 0.0063	0.979 ± 0.0095
CNE (DP)	0.882 ± 0.0045	0.010 ± 0.0039	0.151 ± 0.0080	0.577 ± 0.0058	0.765 ± 0.0284
CNE (EO)	0.959 ± 0.0022	0.064 ± 0.0027	0.043 ± 0.0105	0.651 ± 0.0094	0.971 ± 0.0102
DEBAYES	0.937 ± 0.0026	0.012 ± 0.0042	0.065 ± 0.0074	0.615 ± 0.0067	0.699 ± 0.0524
DOT-PRODUCT	0.895 ± 0.0035	0.068 ± 0.0021	0.149 ± 0.0056	0.738 ± 0.0070	0.975 ± 0.0062
DOT-PRODUCT (DP)	0.745 ± 0.0046	0.003 ± 0.0014	0.032 ± 0.0016	0.570 ± 0.0083	0.500 ± 0.0000
DOT-PRODUCT (EO)	0.892 ± 0.0037	0.043 ± 0.0016	0.043 ± 0.0034	0.657 ± 0.0076	0.946 ± 0.0069
GAE	0.891 ± 0.0031	0.118 ± 0.0035	0.346 ± 0.0309	0.753 ± 0.0138	0.983 ± 0.0061
GAE (DP)	0.775 ± 0.0091	0.002 ± 0.0010	0.031 ± 0.0100	0.583 ± 0.0143	0.825 ± 0.0419
GAE (EO)	0.865 ± 0.0135	0.014 ± 0.0056	0.014 ± 0.0058	0.648 ± 0.0194	0.972 ± 0.0061
MAXENT	0.927 ± 0.0026	0.004 ± 0.0014	0.036 ± 0.0061	0.617 ± 0.0070	/
MAXENT (DP)	0.925 ± 0.0028	0.004 ± 0.0015	0.033 ± 0.0065	0.617 ± 0.0084	/
MAXENT (EO)	0.925 ± 0.0038	0.005 ± 0.0014	0.009 ± 0.0047	0.604 ± 0.0042	/

Table 2. Mean ± standard deviation of results on the ML100K dataset.

METHOD	AUC	DP	EO	RDP	RB
CFC	0.916 ± 0.0021	0.045 ± 0.0086	0.021 ± 0.0070	0.531 ± 0.0040	0.631 ± 0.0191
CFC ($\lambda = 10$)	0.919 ± 0.0021	0.043 ± 0.0065	0.018 ± 0.0045	0.530 ± 0.0030	0.662 ± 0.0127
CFC ($\lambda = 100$)	0.915 ± 0.0023	0.041 ± 0.0069	0.022 ± 0.0066	0.530 ± 0.0036	0.681 ± 0.0158
CFC ($\lambda = 1000$)	0.772 ± 0.0311	0.022 ± 0.0092	0.040 ± 0.0085	0.519 ± 0.0035	0.547 ± 0.0235
CNE	0.906 ± 0.0015	0.046 ± 0.0073	0.119 ± 0.0078	0.536 ± 0.0030	0.547 ± 0.0123
CNE (DP)	0.915 ± 0.0013	0.013 ± 0.0035	0.081 ± 0.0117	0.511 ± 0.0020	0.588 ± 0.0205
CNE (EO)	0.913 ± 0.0014	0.028 ± 0.0055	0.071 ± 0.0124	0.528 ± 0.0032	0.614 ± 0.0134
DEBAYES	0.921 ± 0.0013	0.062 ± 0.0086	0.122 ± 0.0096	0.536 ± 0.0028	0.524 ± 0.0235
DOT-PRODUCT	0.902 ± 0.0012	0.024 ± 0.0051	0.035 ± 0.0033	0.528 ± 0.0029	0.657 ± 0.0233
DOT-PRODUCT (DP)	0.899 ± 0.0012	0.034 ± 0.0058	0.028 ± 0.0026	0.509 ± 0.0037	0.759 ± 0.0232
DOT-PRODUCT (EO)	0.901 ± 0.0012	0.019 ± 0.0026	0.010 ± 0.0045	0.526 ± 0.0029	0.697 ± 0.0275
GAE	0.836 ± 0.0072	0.029 ± 0.0092	0.049 ± 0.0067	0.518 ± 0.0040	0.582 ± 0.0279
GAE (DP)	0.846 ± 0.0112	0.041 ± 0.0140	0.054 ± 0.0094	0.510 ± 0.0050	0.645 ± 0.0195
GAE (EO)	0.866 ± 0.0167	0.030 ± 0.0100	0.030 ± 0.0103	0.517 ± 0.0034	0.578 ± 0.0325
MAXENT	0.902 ± 0.0014	0.008 ± 0.0014	0.042 ± 0.0040	0.536 ± 0.0026	/
MAXENT (DP)	0.885 ± 0.0039	0.003 ± 0.0011	0.029 ± 0.0065	0.515 ± 0.0055	/
MAXENT (EO)	0.899 ± 0.0013	0.006 ± 0.0009	0.012 ± 0.0030	0.530 ± 0.0018	/

Table 3. Mean ± standard deviation of results on the FACEBOOK dataset.

METHOD	AUC	DP	EO	RDP	RB
CFC	0.981 ± 0.0015	0.009 ± 0.0050	0.015 ± 0.0084	0.529 ± 0.0068	0.601 ± 0.0150
CFC ($\lambda = 10$)	0.984 ± 0.0014	0.009 ± 0.0039	0.016 ± 0.0058	0.529 ± 0.0040	0.615 ± 0.0153
CFC ($\lambda = 100$)	0.985 ± 0.0015	0.008 ± 0.0041	0.015 ± 0.0038	0.529 ± 0.0049	0.617 ± 0.0157
CFC ($\lambda = 1000$)	0.904 ± 0.0381	0.011 ± 0.0038	0.024 ± 0.0071	0.532 ± 0.0052	0.586 ± 0.0228
CNE	0.982 ± 0.0006	0.001 ± 0.0007	0.084 ± 0.0050	0.538 ± 0.0035	0.601 ± 0.0189
CNE (DP)	0.976 ± 0.0006	0.004 ± 0.0010	0.031 ± 0.0030	0.512 ± 0.0027	0.594 ± 0.0217
CNE (EO)	0.980 ± 0.0006	0.004 ± 0.0010	0.006 ± 0.0035	0.518 ± 0.0025	0.607 ± 0.0192
DEBAYES	0.981 ± 0.0004	0.001 ± 0.0007	0.087 ± 0.0042	0.539 ± 0.0030	0.592 ± 0.0163
DOT-PRODUCT	0.989 ± 0.0006	0.002 ± 0.0005	0.015 ± 0.0007	0.561 ± 0.0037	0.630 ± 0.0138
DOT-PRODUCT (DP)	0.987 ± 0.0008	0.001 ± 0.0008	0.003 ± 0.0007	0.515 ± 0.0032	0.500 ± 0.0000
DOT-PRODUCT (EO)	0.989 ± 0.0005	0.003 ± 0.0007	0.005 ± 0.0008	0.530 ± 0.0041	0.769 ± 0.0185
GAE	0.981 ± 0.0028	0.004 ± 0.0016	0.031 ± 0.0050	0.545 ± 0.0037	0.602 ± 0.0152
GAE (DP)	0.977 ± 0.0027	0.003 ± 0.0017	0.016 ± 0.0039	0.518 ± 0.0061	0.626 ± 0.0193
GAE (EO)	0.974 ± 0.0042	0.004 ± 0.0016	0.005 ± 0.0020	0.535 ± 0.0026	0.618 ± 0.0240
MAXENT	0.845 ± 0.0014	0.001 ± 0.0003	0.012 ± 0.0007	0.541 ± 0.0041	/
MAXENT (DP)	0.838 ± 0.0043	0.000 ± 0.0002	0.007 ± 0.0012	0.530 ± 0.0051	/
MAXENT (EO)	0.824 ± 0.0180	0.001 ± 0.0006	0.002 ± 0.0016	0.519 ± 0.0070	/

References

1. Bose, A., Hamilton, W.: Compositional fairness constraints for graph embeddings. In: International Conference on Machine Learning. pp. 715–724 (2019)
2. Buyl, M., De Bie, T.: Debayes: a bayesian method for debiasing network embeddings. In: International Conference on Machine Learning. pp. 1220–1229. PMLR (2020)
3. Kallus, N., Zhou, A.: The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. Advances in neural information processing systems **32** (2019)