ID2211 — Link Prediction on MovieLens with regard to Fairness

Andri Furrer Ankit Grover Senne Vanden Eynde Titouan Mazier

June 10, 2024

1 Introduction

In recent years the focus has shifted from solely increasing accuracy to also incorporating fairness in recommender systems, to tackle the problem of unfair recommendations. Unfair recommendations are problematic when recommendation systems are deployed to billions of users in social networks, streaming platforms, and E-commerce websites. Recommender systems perform link prediction, which is the problem of predicting if a link between two nodes in a network exists, where a node can be any network entity. Regarding fairness, many aspects can be considered: is it fair for an individual? A group? Or is it fair for providers? Examples of group fairness issues are a lack of represented social groups, job opportunities that are not advertised fairly to everyone, or the creation of filter bubbles.

A good example of link prediction and fairness combined would be recommender systems for movies. Because movies play an important part in the construction of one's identity, it is important not to reproduce existing segregation in the recommendations. We don't want to see men being recommended action movies and women being recommended children's movies simply because these genres are more popular among people of the same gender. It may be fairer if both get a science-fiction movie for their next recommendation for example. In general, recommendations should be based on their personal preferences and not on group preferences.

We define our research question as *Do network-based movie recommendation systems show a gender bias with respect to movie genres?*. In the research project, we evaluated link prediction methods on the MovieLens-100K dataset regarding fairness. We first evaluated the accuracy of classical methods on the given dataset, followed by an evaluation of state-of-the-art Graph Neural Networks on our dataset. This gives us insight into how network topology as well as features influence fairness in movie recommenders. Our findings suggest that all models exhibit small biases with respect to gender.

2 Related Work

The link prediction problem is well established. The first solution proposed to solve it was to measure similarities between each pair of nodes and use these similarities as an indicator of how likely it is for an edge to form between these two nodes. Such similarities include neighborhood-based measures such as (among others) common neighbors[1], Adamic-Adar similarity[2] or preferential attachment[1]. Other methods use Random Walk probabilities, two of them are rooted PageRank[1] and PropFlow[3]. Later Lichtenwalter et al. proposed to approach link prediction as a supervised node-pairs clustering problem[3] and achieved improved results and complexity. This approach has later been refined using Neural Networks for both clustering and feature generation with the arrival of the Graph Neural Network (GNN)[4].

However, these methods used to be evaluated solely based on prediction accuracy and recall. In the late 2010s concerns about the fairness of the solutions started to appear. In their exhaustive study, Dong et al. present a taxonomy for the different fairness problems that can be addressed in link prediction[5]. One of these is *Individual fairness* and can be summarized as the following: *Do similar nodes get similar predictions?*. This fairness criterion often collides with *Group Fairness* that tries to evade discriminatory predictions toward users from the same sensitive subgroup (who share a common *sensitive feature*).

Group unfairness is usually caused by the relay or amplification of a bias in the initial dataset, thus one solution proposed to improve group fairness consists of pre-processing the graph on which link prediction is performed to reduce the bias contained in it[6, 7]. Similarly, it is possible to post-process the results of the link prediction to mitigate the bias on the prediction[8, 9]. What seems to be the cutting-edge solution, however, is to use an adversarial learning network to try to maximize the accuracy of the results while making sure that group fairness is respected[10, 9].

To evaluate these models based on fairness, previous works used (among others) Demographic Parity/Equality of Odds[10, 6], Statistical Parity, Equality of Opportunity or modularity reduction[9, 8]. The following paragraphs discuss these concepts in more detail.

The demographic parity criteria is defined by Dong et al. as[5]:

$$P(\hat{Y} = 1|S = 0) = P(\hat{Y} = 1|S = 1)$$

Where $\hat{Y} = 1$ is a positive edge prediction and S indicates a sensitive pair of nodes (two nodes whose sensitive feature "match" in some unwanted way). This criterion is then developed into numerical evaluations of how close a given model lands from achieving it.

Equality of Odds adds to this concept by taking into account the ground truth edges in the evaluation. Dong et al. define it as verifying the following conditions[5]:

$$\begin{cases} P(\hat{Y} = 1|S = 0, Y = 0) = P(\hat{Y} = 1|S = 1, Y = 0) & \text{True positive parity} \\ P(\hat{Y} = 1|S = 0, Y = 1) = P(\hat{Y} = 1|S = 1, Y = 1) & \text{False positive parity} \end{cases}$$

Where $Y = 1$ indicates the presence of an edge in the ground truth data.

Modularity reduction measures the evolution of the graph modularity implied by the link prediction process.

3 Data and Experimental setup

We work with the MovieLens 100k dataset which provides a list of movie ratings performed by users. For each movie, we have access to its name, release date, and a set of 18 genres, with each movie belonging to any number of genres. The dataset also provides the user's gender, age, and occupation. Each rating is a single value among $\{1, 2, 3, 4, 5\}$ possible stars, 1 meaning that the user did not like the movie and 5 that they loved it. Table 1 provides a summary of the dataset. Our goal will be to provide each user with movies they will likely rate positively.

Number of users	943
Number of movies	1,682
Number of ratings	100,000
Number of 4&5 star ratings	55,375
Average number of 4&5 ratings by user	58.7
Average number of 4&5 ratings by movie	32.9

Table 1: Statistics of the MovieLens Dataset 100k.

Type of graph	Bipartite
Average shortest path	2.9
Diameter	6

Table 2: Networks characteristics of the graph created from the MovieLens 100k Dataset.

To simplify the problem we only keep the 4 and 5-star ratings, that is, assigning a 1 to the 4 and 5-star ratings, and a 0 to the 1-3-star ratings, effectively reducing the problem to a binary one. We then have links between our users if they rated a movie with a positive 4-5 star rating or 0 otherwise. With that, we only have 55k edges in our graph left.

We then map the ratings to a bipartite graph where each positive rating appears as an edge between the rated movie and the user who submitted the rating.

The focus of this work was on group fairness. Our first idea was to question the relationship between users age and the release date of the movie they rated, however it seems that such a bias is very unlikely both as a result of the recommendation and as a ground truth in the real-world tendencies. This brought us to the research question stated in the introduction: *Do network-based movie recommendation systems show a gender bias with respect to movie genres?*. The first step was to analyze if such a bias was naturally present in the dataset. Our assumption relies on the fact that the majority of the bias is introduced through the data and then exaggerated by the algorithm's characteristics to an extent.

Figure 3 shows the distribution of user's 4-5-star ratings of the different genres present in the datasets. We can observe two things. Firstly, there is an important bias in the overall dataset with a large majority of the users being men. Secondly, the relative bias in rating is quite uniform with a difference between women and men comprised between -30% and +40% for every movie genre. Before applying our methods, we split our dataset into training and test sets. We use 80 % of the data in the training set and 20 % of the data in the test set. For the GNN, we use random seeds to split our data into batches for training the model.

The fairness metrics we use are *Statistical Parity* (SP) and *Equality of Opportunity* (EO). Statistical Parity (SP) assesses whether the proportion of positive outcomes is the same for all groups:

$$SP = P(\hat{Y} = 1 | G = Female) - P(\hat{Y} = 1 | G = Male) \tag{1}$$

Equality of Opportunity (EO) assesses whether the probability of a positive outcome is the same for individuals from different groups who have the same qualifications:

$$EO = P(\hat{Y} = 1 | Y = 1, G = Female) - P(\hat{Y} = 1 | Y = 1, G = Male) \tag{2}$$

4 Methods

In this section, we present the method used for link prediction. These methods do not intend to improve the fairness of the result.

4.1 Unsupervised approaches

First, we tried to implement unsupervised methods listed by Liben-Nowell and Kleinberg[1] based on node neighborhood (Common neighbors, Adamic-Adar, Jaccard coefficient and preferential attachment, see [1]). Unfortunately, except for preferential attachment, these metrics do not work for user-item recommendation on a bipartite graph as all the neighbors of a user will be items and all the neighbors of an item will be users. There is therefore no common neighborhood for a user and an item that we can measure. An adaptation is to look just a little further: instead of measuring the intersection of the neighborhoods, we look at the edge set between the two (distinct) neighborhoods $I(\mathcal{N}(x), \mathcal{N}(y)) = \{(u, v) \in E | u \in \mathcal{N}(x), v \in \mathcal{N}(y)\}$ (see Figure 1).

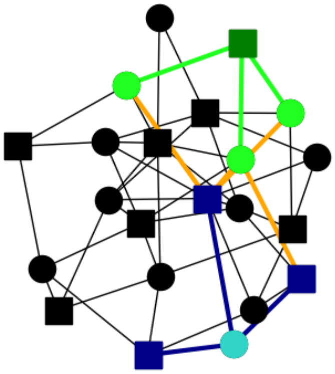


Figure 1: Illustration of $I(\mathcal{N}(x),\mathcal{N}(y))$ (in orange). x and y are respectively the light blue and dark green nodes while $\mathcal{N}(x)$ and $\mathcal{N}(y)$ are respectively the dark blue and light green nodes.

Table 3 details the metrics adapted to bipartite graphs with this method.

As the fairness measurements of the models are done on a static prediction, we have to select a threshold to split the ranking of the potential edges given by the models into a list of positive/negative predictions. We do so by selecting as many predicted edges as the number of edges in the base dataset.

Similarity Metric	**Original definition**	**Adaptation to bipartite graphs**
Common neighbors	$\lvert\mathcal{N}(x)\cap\mathcal{N}(y)\rvert$	$\lvert I(\mathcal{N}(x),\mathcal{N}(y))\rvert$
Adamic-Adar	$\sum_{z\in\mathcal{N}(x)\cap\mathcal{N}(y)}\frac{1}{\log(d(z))}$	$\sum_{(u,v)\in I(\mathcal{N}(x),\mathcal{N}(y))}\frac{1}{\log(d(u)+d(v))}$ *
Jaccard's coefficient	$\frac{\lvert\mathcal{N}(x)\cap\mathcal{N}(y)\rvert}{\lvert\mathcal{N}(x)\cup\mathcal{N}(y)\rvert}$	$\frac{\lvert I(\mathcal{N}(x),\mathcal{N}(y))\rvert}{\lvert\delta(\mathcal{N}(x))\cup\delta(\mathcal{N}(y))\rvert}$ **
Preferential attachment	$d(x)\cdot d(y)$	$d(x)\cdot d(y)$

* $d(u)$ is the degree of node u.

** $\delta(S)$ is the edge boundary of the node subset S.

Table 3: Definition of the bipartite adaptations of the similarity metrics.

4.2 Graph Neural Networks

As a second approach, we used Graph Neural Networks as a recommendation system. We considered user and movie nodes as heterogeneous to utilize rich and distinct node-level feature information besides topology. Edges were made bi-directional specifically to facilitate message-passing process for training the GNN. Since training an entire graph at once can be computationally expensive, we sample sub-graphs for training. The sample are created by sampling a few edges from the edge links provided (including negative edges) and then adding 1-hop and 2-hop neighbors of these edge's extremities to the sub-graph. The architectures we experimented with include `GATConv` [11], `GINConv`[12], `GraphConv`[13], and `SAGEConv`[14] which are compatible with bipartite graphs[15]. The general architectural setup of our classifier can be seen in Figure 2. This corresponds to a homogeneous GNN architecture before conversion to a more complex heterogeneous one. The GNN takes node embeddings as well as feature projections and then predicts logits that are passed through `Softmax` to get probabilities. `BinaryCrossEntropy` was used as a loss function with an `Adam` optimizer[16]. We hypothesize that these architectures would be good at capturing local network information with respect to user-movie interactions.

To get an idea of what data features were best to evaluate for fairness, we performed an ablation study. The focus of the ablation study was to see what user and movie features would be utilized for training the GNN architectures. This allows for a fairer evaluation after fixing the features and noticing the change in the evaluation and fairness metrics. We chose a baseline model such as `SAGEConv` and tried different user feature combinations. The idea was to choose the feature combinations that

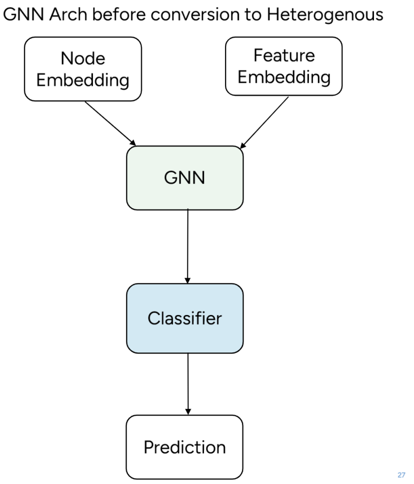


Figure 2: General architecture of the GNN, with a classification head.

provide the best results in terms of both accuracy and fairness metrics. This was done to minimize fairness issues due to the absence of fairness-aware training.

Then, experiments were performed to get a better understanding of how the different architectures might influence the fairness scores. We checked if a correlation could be found between the fairness metrics and accuracy metrics received from the models and then focused on the results of the well-performing models (with regard to accuracy and fairness), examining if any induced biases of fairness could be found. For this, we zoomed in on the movie genres and looked at the gender proportions in the predictions from the models and the raw data from the dataset. Finally, we calculated the fairness metrics with regard to the movie genres on the predictions of the models, in the hope of gaining a better understanding of any potentially induced biases.

5 Results

In the following section, we present the results of the ablation study, the accuracy and the fairness evaluation of the models presented in the last section.

5.1 Ablation Study

The results of the ablation study are showcased in table 4.

Feature Ablation	**AUC**	**Accuracy**	**F1**	**SP**	**EO**
Age + Occ + Gender, Movie	**0.8643**	**0.7998**	0.6809	0.036829	0.047949
Age + Occ, Movie	0.8635	0.7971	**0.6837**	0.026296	0.030179
Age, Movie	0.8061	0.7310	0.4290	**0.020913**	**0.021649**

Table 4: Performance metrics for different feature ablations.

We clearly see a tradeoff between accuracy and fairness here. The combination of Age, Occupation, and Movie features provides the best balance in terms of performance and fairness metrics. Although removing occupation and gender features does increase fairness, probably due to implicitly learned gender information from occupation, they heavily affect accuracy. Hence, these features were used for evaluation based on accuracy and fairness of all architectures.

Model	AUC	Accuracy	F1-score	Precision	Recall
Bi-common-neighbors	0.8584	0.7424	0.7001	0.5888	0.8632
Bi-Adamic-Adar	0.8599	0.7436	0.7017	**0.5907**	0.8642
Bi-Jaccard	**0.8695**	**0.7448**	**0.7027**	0.5905	**0.8674**
Preferential Attachment	0.8450	0.7273	0.6781	0.5653	0.8472

Architecture	Layers	AUC	Accuracy	F1-score	Precision	Recall	SP	EO
GATConv	3	0.8013	0.7439	0.5956	0.5656	0.6288	0.007911	**0.007694**
	4	0.8306	0.7609	0.6587	0.6923	0.6283	0.009295	0.008364
	5	**0.9215**	**0.8504**	**0.7801**	**0.7963**	0.7647	0.024848	0.018533
	6	0.7706	0.6751	0.0635	0.0331	**0.8097**	**0.004536**	0.013637
GINConv	3	0.7118	0.7025	0.5692	0.5895	0.5502	0.026767	0.001744
	4	**0.8606**	**0.8041**	**0.7365**	0.8214	**0.6675**	0.016521	0.006719
	5	0.7960	0.5924	0.5766	**0.8324**	0.4410	**0.004849**	**0.000996**
	6	0.5927	0.6921	0.5363	0.5342	0.5384	0.012549	0.004277
GraphConv	3	0.8822	0.8142	0.7047	0.6650	0.7494	0.018200	0.006651
	4	0.8853	0.8144	0.6977	0.6426	**0.7632**	0.021984	0.023273
	5	0.8822	0.8130	0.7288	0.7538	0.7053	**0.011697**	0.005238
	6	**0.8914**	**0.8187**	**0.7355**	**0.7564**	0.7158	0.030041	**0.000958**
SAGEConv	3	0.8767	0.8100	0.6994	0.6632	0.7399	**0.009929**	0.016642
	4	0.8782	**0.8116**	**0.7193**	0.7244	0.7143	0.028708	0.008952
	5	**0.8804**	0.8101	0.6992	0.6621	**0.7406**	0.016706	0.003372
	6	0.8747	0.8009	0.7135	**0.7439**	0.6855	0.017205	**0.001091**

Table 5: Performance measures for the different models tested.

5.2 Accuracy

Table 5 lists the performance of the different models. We can see that unsupervised methods provide relatable results despite their conceptual simplicity. We can note that Preferential Attachment is slightly less performant, this loss of performance is however considerably balanced by a way better time complexity. The GNN models provide even better results, especially for binary predictions (evaluated by Accuracy, F1, Precision, and Recall scores).

5.3 Fairness

Figure 3 presents the bias between movie genres and user gender for the base data, the best-performing unsupervised model (Bi-Jaccard), and the best-performing GNN (`GATConv` with 5 layers). We can see that the overall bias towards men is still present in the predictions. However, the bias toward specific genres appears more flattened out, sometimes even counterbalanced, by the predictions.

In Figure 5 the fairness scores of the models are plotted against different relevant parameters or metrics of the different GNN tested. There is no visible correlation between the amount of parameters and the fairness it achieves. The way the overall fairness changes, with different parameters, is not significant, as can be seen by the scale of the y-axis. Likewise, for accuracy and F1-score, we don't see a significant causal-effect despite slight correlation between evaluation metrics and the fairness metrics. This is obvious since training without a fairness objective would optimize for accuracy making fairness metrics worse.In general, the fairness values do get worse, but only by a small magnitude.

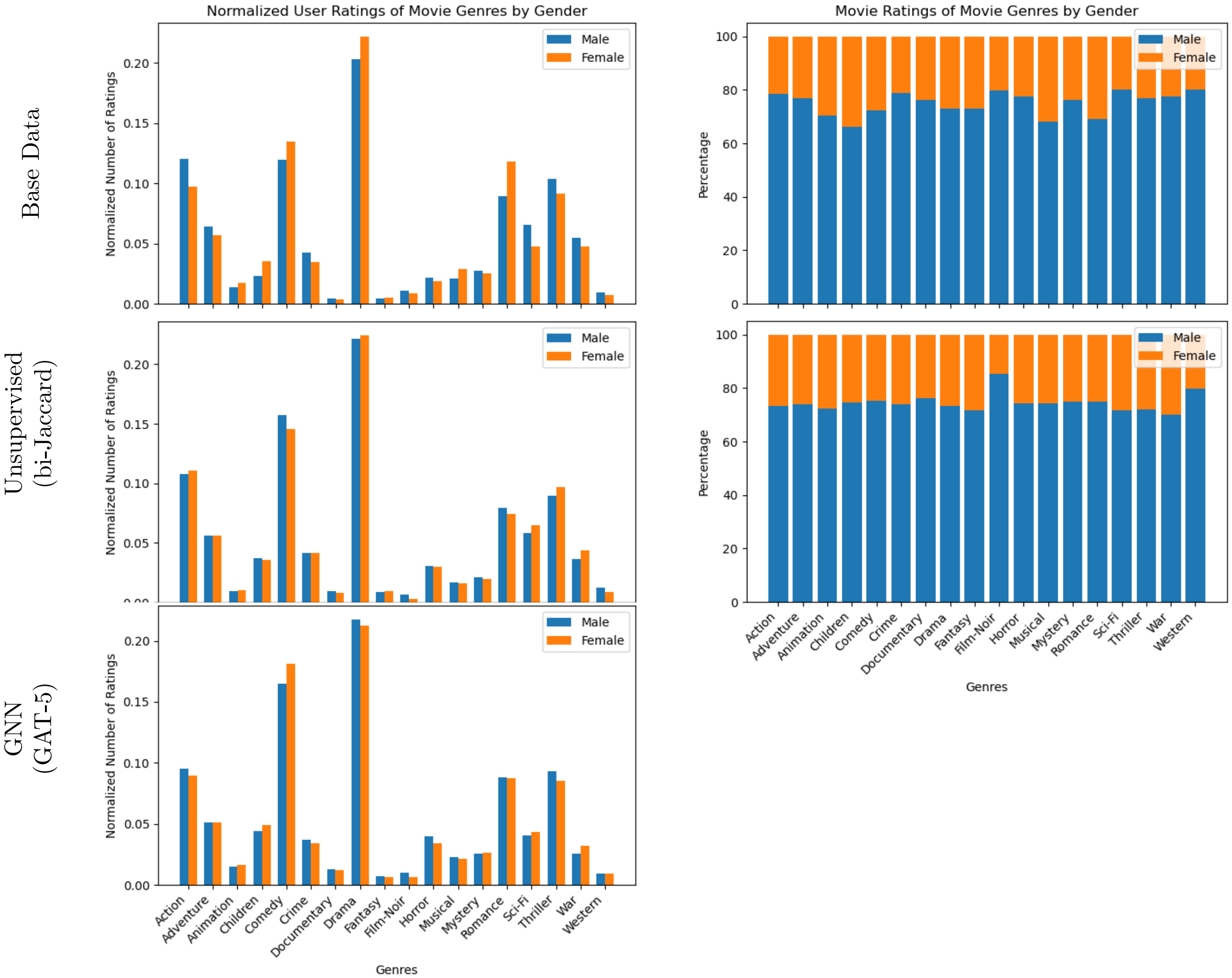


Figure 3: Distribution of the 4+ ratings based on user gender and movie genre. The models tested do not seem to introduce any new bias. They even slightly counterbalance the initial biases in some cases.

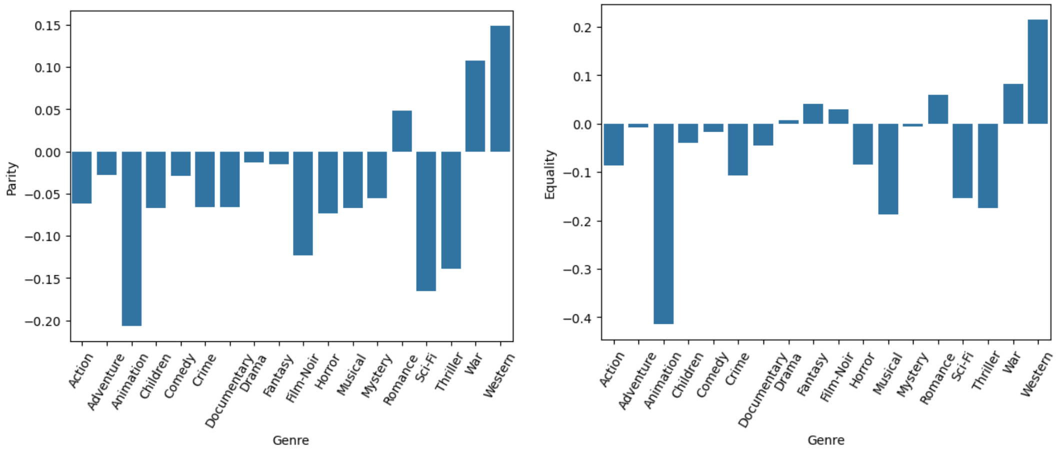
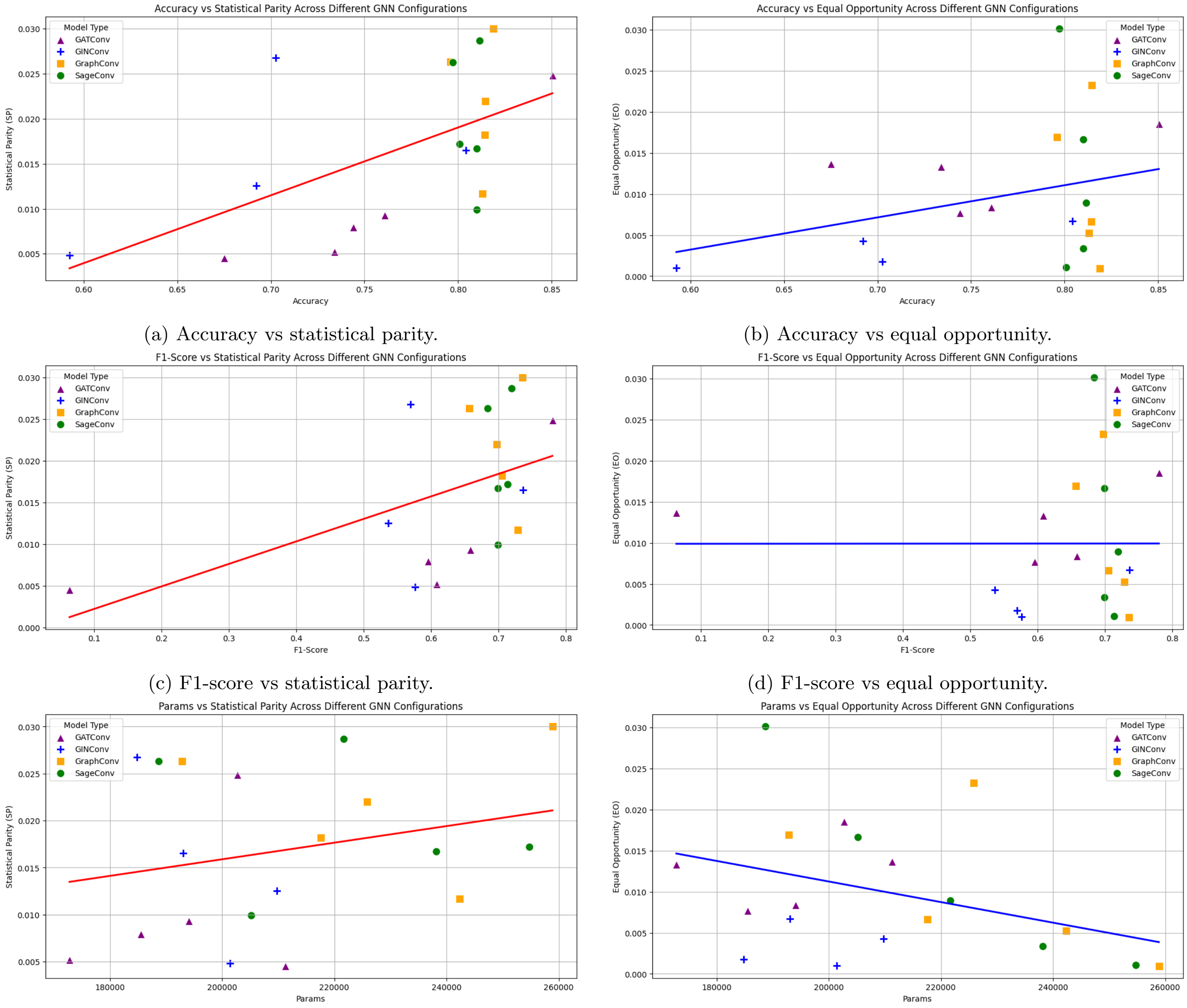


Figure 4: Parity (SP) and Equality (of opportunity, EO) measures calculated for every movie genre for `GATConv-5L`. A positive sign indicates a bias towards female users. A negative sign means a bias towards male users.



(a) Accuracy vs statistical parity.

(b) Accuracy vs equal opportunity.

(c) F1-score vs statistical parity.

(d) F1-score vs equal opportunity.

(e) Amount of parameters vs statistical parity.

(f) Amount of parameters vs equal opportunity.

Figure 5: Relation between accuracy measures, network parameters, and fairness measures for the different GNN architectures. The fairness metrics seem to be correlated, however, no causal effect can be established. Correlation of fairness with network complexity is extremely low suggesting almost no impact. Also, statistical parity and equal opportunity vary only slightly among the different models.

`GraphConv` with 3 layers, and `GATConv` with 5 layers seem to perform best when considering both accuracy and fairness metrics. Hence, we will zoom in a bit more on one of these models to get some insights into any biases they might have. For this discussion, we will focus on the `GATConv` model (denoted as `GATConv-5L`). Figure 3 presents the genre-specific biases measured on the prediction made by the model. The model does not significantly alter the distribution of the dataset for the predictions – the proportions across categories remain more or less the same for the same gender. However, it does change the disparity between the genders across for some categories. This indicates there is some sort of induced bias present, albeit small.

If we calculate the fairness metrics for the predictions with regard to the movie genres, rather than all the features, we get the following plots as seen in Figure 4. The statistical parity and equality of opportunity are plotted but without the absolute value signs usually seen for these formulas. This way one can observe in what direction the category is biased: a positive value means it is biased towards female users, and a negative value means it is biased towards male users.

The `GATConv-5L` model heavily favors male users for most categories. We noticed a similar phenomenon with all the other models. We think this is because of the disproportion of 4-5-star reviews left by the different genders in the dataset: there are significantly more male users leaving positive reviews than female users. We notice the models implicitly learn the gender rating distribution and introduce more links toward male users in that manner. Therefore, the recommendation system we consider here does introduce a bias towards male users.

5.4 Interpretation

What does it mean to have a bias toward male users? The static prediction generated by the model is a set of edges that were not present in the training set but are likely to appear in the test set. i.e. it is a set of ratings that are likely to be made by users in the future. A bias in this set toward male users means that a majority of these ratings are expected to be made by male users. This is of course due to the imbalance between female and male users in the dataset, however, the bias still appears with the equality of opportunity that is designed to address such imbalance, which means that there is more to explain it.

In reality, we expect the rate of female/male users in the future ratings to be strongly independent of the network structure. Thus, the bias observed is unlikely to impact the real future of the graph. What it does, however, is skew the predictions toward male users, with two consequences: biasing any measurement performed on predicted data and give worse predictions to female users whose true positive rate is less represented in the model validation.

6 Conclusion

In this project, we showed that we could adapt simple neighborhood-based unsupervised methods for link prediction to bipartite graphs by slightly extending the range of the aforementioned "neighborhood". These new similarities perform well on the MovieLens100k dataset with the Bi-Jaccard method performing best for AUC, accuracy, F1-score, and recall, only outperformed by Bi-Adamic-Adar for precision.

We also evaluated different GNN architectures and achieved even better results in terms of both fairness and performance. The `GATConv` model performed best for AUC, accuracy, F1-score, recall and statistical parity whereas `GINConv` did best for precision and `GraphConv` for equality of opportunity. Hence, we can conclude that `GATConv` is a good model to choose from, if one wants to have good accuracy and decent fairness properties.

We demonstrated that network-based link prediction methods barely introduce or exacerbate group bias in the prediction, as the distribution of predicted ratings on each specific genre tends to balance out gender imbalance. However, the models still introduce a global bias toward male users with consequences on the prediction validity for further data mining use and expected worse prediction for female users.

7 Further Research Questions

The inherent bias due to the dataset conception is unknown, so a further research project could go in that direction. We also have not conducted training of models using fairness-aware metrics, this could provide some interesting findings. Reweighing the outputs or rebalancing the number of edges based on gender as well as probing into GNN representations for fine-grained node-level explanations could provide promising results.
It is also worth noting that this project focused on finding bias toward movie genres. One might get different results while looking for bias toward other categorizations of movies.

8 Detail of the Contributions

The work on this project was split along two main lines: the analysis of the dataset and the unsupervised methods performed by Andri F. and Titouan M. and the training and evaluation of GNNs performed by Ankit G. and Senne V.E. We then all worked together to write the report.

References

[1] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[2] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.

[3] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, (New York, NY, USA), p. 243–252, Association for Computing Machinery, 2010.

[4] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), p. 855–864, Association for Computing Machinery, 2016.

[5] Y. Dong, J. Ma, S. Wang, C. Chen, and J. Li, "Fairness in graph mining: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 10, pp. 10583–10602, 2023.

[6] O. D. Kose and Y. Shen, "Fairness-aware adaptive network link prediction," in *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 677–681, 2022.

[7] C. Laclau, I. Redko, M. Choudhary, and C. Largeron, "All of the fairness for edge prediction with optimal transport," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* (A. Banerjee and K. Fukumizu, eds.), vol. 130 of *Proceedings of Machine Learning Research*, pp. 1774–1782, PMLR, 13–15 Apr 2021.

[8] A. Saxena, G. Fletcher, and M. Pechenizkiy, "Hm-eiict: Fairness-aware link prediction in complex networks using community information," *Journal of Combinatorial Optimization*, vol. 44, pp. 2853–2870, Nov 2022.

[9] F. Masrour, T. Wilson, H. Yan, P.-N. Tan, and A. Esfahanian, "Bursting the filter bubble: Fairness-aware network link prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 841–848, Apr. 2020.

[10] M. Cao, J. Song, J. Yuan, B. Zhang, and C. Wang, "Fairhelp: Fairness-aware heterogeneous information network embedding for link prediction," in *Database Systems for Advanced Applications* (X. Wang, M. L. Sapino, W.-S. Han, A. El Abbadi, G. Dobbie, Z. Feng, Y. Shao, and H. Yin, eds.), (Cham), pp. 320–330, Springer Nature Switzerland, 2023.

[11] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[12] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," in *International Conference on Learning Representations*, 2019.

[13] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe, "Weisfeiler and leman go neural: Higher-order graph neural networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4602–4609, Jul. 2019.

[14] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[15] "GNN Cheatsheet &#x2014; pytorch_geometric documentation — pytorch-geometric.readthedocs.io." https://pytorch-geometric.readthedocs.io/en/2.5.0/cheatsheet/gnn_cheatsheet.html. [Accessed 07-06-2024].

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

Appendix A Network analysis

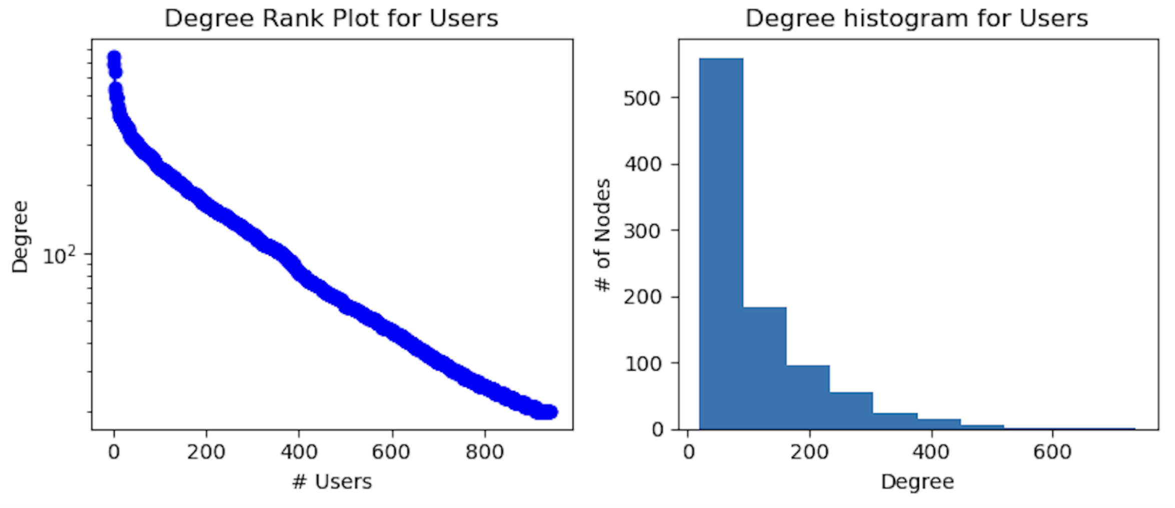


Figure 6: The degree distribution of the user degrees follows a power law.

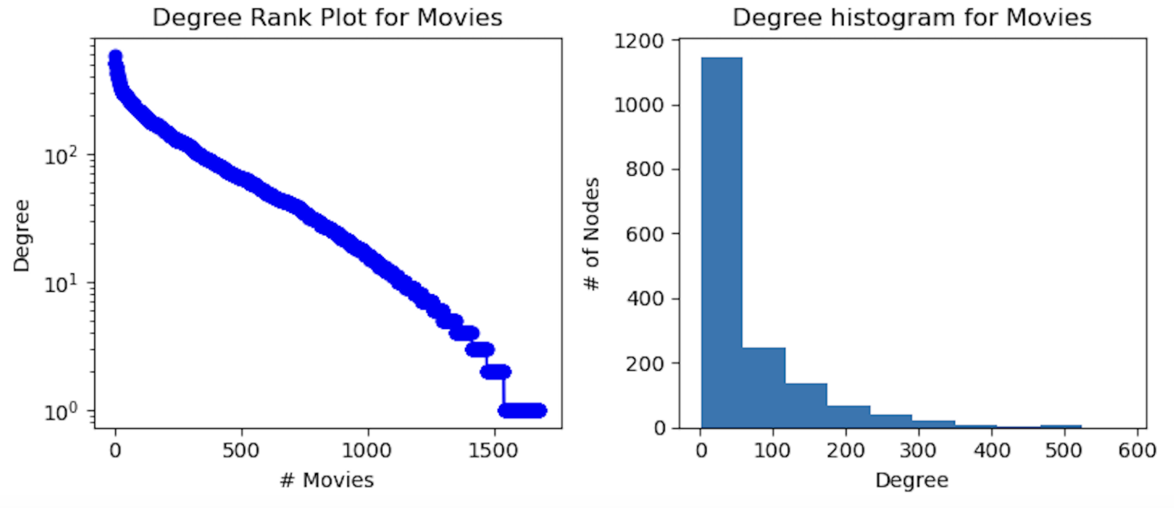


Figure 7: The degree distribution of the movie degrees follows a power law.

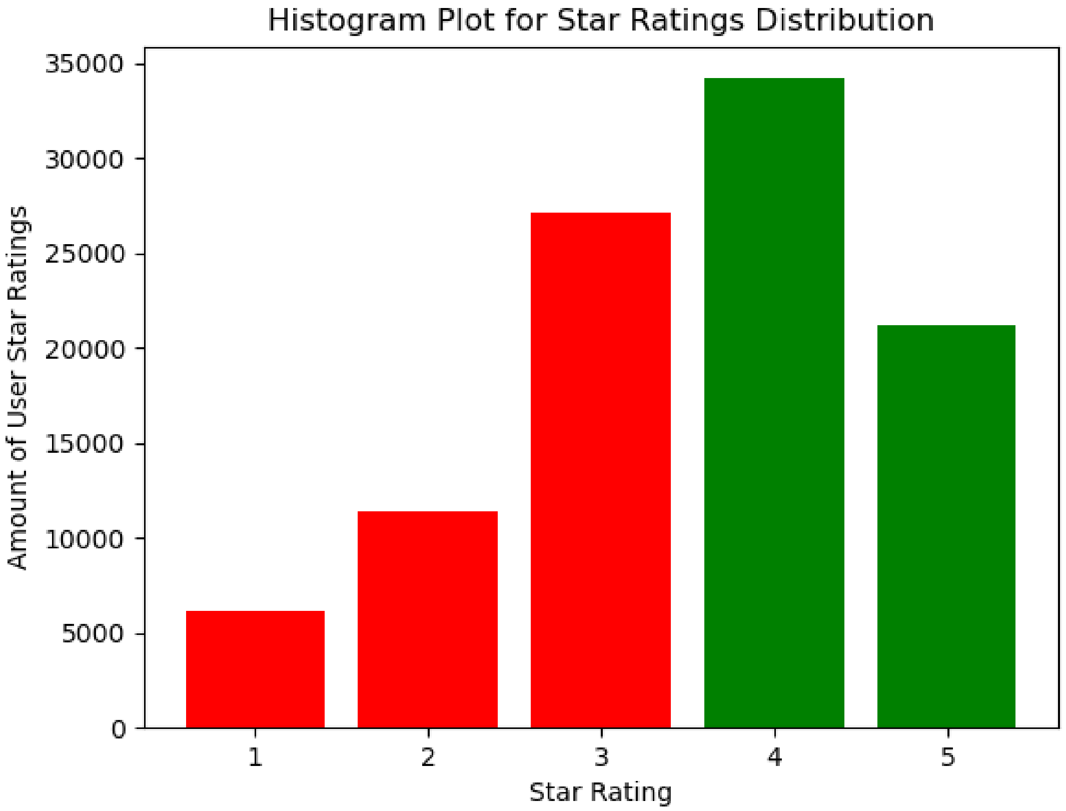


Figure 8: The 1- to 5-star ratings distribution, where the positively considered 4- and 5-star ratings are colored green and the negatively considered 1-3-star ratings are colored red.