Enhancing Recommendation Diversity in Trust-Based GNNs Through Adversarial Regularisation

Dhruvi Viral Goradia[1], Kaustubh Girish Sharma[2], Akshata Pravin More[3], Abhishek Vichare[4]

{dhruvi.goradia18@nmims.in[1], kaustubh.sharma47@nmims.in[2], akshata.more03@nmims.in[3]}

Department of Computer Engineering, Mukesh Patel School of Technology Management & Engineering, SVKM's NMIMS Deemed-to-be University, Mumbai, 400056, India[1,2,3,4]

Abstract. Although the recommender systems based on social trust may be correct, a major issue with them is that they tend to increase popularity bias and form an echo chamber due to their strong tendency to recommend mainstream content in the first place. In this paper, the audit of a baseline trust-aware Graph Neural Network (GNN) is used to measure this bias and a new framework to reduce it is proposed. We suggest an Adversarial regularisation mechanism, which re-formulates the training process into a competitive game. In this type of system, the GNN recommender will have to race an opponent network, which we define as the Discriminator. This Discriminator is trained to identify the biased patterns of the recommenders, which will force them to come up with more diverse and less predictable recommendations. Our empirical findings confirm the effectiveness of the approach, as there was an absolute variation of twenty percent in Popularity Bias that reduced in the baseline model and reached 39.90% to 19.90%. Importantly, this remarkable gain in fairness was attained without worsening the predictive performance, indicating a feasible direction in the creation of systems that are highly performant yet foster content discovery.

Keywords: Recommender Systems, Graph Neural Networks, Adversarial Regularisation, Popularity Bias, Trust-Aware Systems.1 Introduction

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2366241

1 Introduction

Recommender systems are important tools that assist us in sifting through a large load of content, such as movies or news. One common approach to enhancing these recommendations is via social trust networks, which are based on the intuitive notion that you are better able to enjoy something that your friend suggests. Although modern technologies (such as Graph Neural Networks (GNNs)) are quite effective in utilising this trust information to achieve accuracy, they conceal a significant danger. These systems, when trained to learn based on the social circle of one user, have the potential to generate and reinforce algorithmic bias, resulting in filter bubbles and a popularity bias where only mainstream content is suggested.

Using social connections (like friend lists) has helped recommender systems make much better predictions, especially when there isn't enough data on a user. However, this improvement comes with a hidden side effect. It makes everyone's recommendations look quite literally the same. This happens because people naturally tend to be friends with others who are similar to them. When the system learns strictly from these friendships alone, it accidentally tends to traps users in a filter bubble. In this bubble, users are only shown the popular and mainstream content that their friends have already liked.

This creates a repetitive cycle. The system suggests popular items, users click on them, and the system becomes even more convinced that it should only show popular items. There is an urgent need for a new method that can use social data to be accurate, and at the same time also break this cycle to help users discover new and more diverse content. This study is motivated by the need to move away from systems that focus only on accuracy, and instead build systems that are responsible and fair in the content that they recommend.

Until recently, much of the research has been on accuracy improvement, with little to say on whether these systems are fair and socially impactful. It is an extreme discrepancy, because biased recommenders can constrain user discovery and form echo chambers.

To address the limitations of existing trust-based models, this paper makes the following key contributions:

- Empirical Bias Audit: We construct a standard Trust-Aware GNN baseline and perform a quantifiable audit, establishing empirical evidence that trust-based aggregation results in a high Popularity Bias (39.90%) and measurable Homogeneity Score (1.14%), confirming the echo chamber hypothesis.
- Novel Adversarial Framework: We propose a fairness-aware architecture that integrates a Discriminator network with the GNN Recommender. Unlike static re-ranking methods, this framework formulates fairness as a dynamic, competitive minimax game where the recommender learns to generate embeddings that are predictive of ratings but invariant to community clustering.
- Demonstrable Fairness Gain: We demonstrate through extensive experimentation on the Epinions dataset that our method achieves a 20.00% absolute reduction in popularity bias and reduces the homogeneity score to 0.81%, with negligible degradation in predictive accuracy (RMSE).

Our results show it is possible to design a system that uses social trust to make accurate recommendations and actively encourages fairness. We conclude that our adversarial model can significantly decrease popularity bias while still retaining a high degree of predictive power, offering a viable way to develop more responsible and useful recommender systems.

2 Related Works

The general assumption of the social recommender systems is that social affiliation is a solid predictor of user preference; in fact, the taste of the friends of a user provides a strong signal on the taste of the user himself/herself. The initial studies in this area proposed a model that synthesised personal preferences, general popularity of the item and social influence [1]. This innovative work showed that the integration of social network information significantly enhances predictive accuracy, thus reducing the constant issues of data sparsity, i.e. low user ratings, and the cold-start problem of new users. In the long run, the history of development of the sphere has been outlined. As an example, recent surveys have highlighted the importance of contextual variables in social networks, in that relationships do not rely solely on the aspect of connectivity [2]. The trend towards deep learning algorithms is traced in other reviews [3], and the joint system of approaches is systematically investigated in the field of optimising the quality of the recommendations, where it is emphasised that the integration of multiple approaches is the present-day practice [4]. Altogether, these studies assume that the successful personalisation requires the understanding of the definite context underlying social interactions.

One of the areas that have received a special interest in current studies is the enhancement of Collaborative Filtering (CF) methods with the integration of social data. It has also been proposed that methodological extensions can be made to canonical CF algorithms that can calculate user similarity through a weighted sum of explicit ratings and social relational information [5]. This change is based on the assumption that the impact of friends of a person is more than that of an unknown user who has similar interests. The most accurate way to quantify such relationships has been determined by having comparative analyses of specific similarity measures, such as PageRank and Jaccard similarity, to identify the best method of quantifying a relationship [6]. In addition, combined community detection mechanisms with CF provided systems have been able to give preference to suggestions with a source within the individual social group of a user [7]. Similar studies have focused on finding the most authoritative or appropriate recommenders among individuals in their social circle to improve the quality of the recommendations, in general, in parallel studies [8].

Going further than elementary connectivity, researchers have explored the granularity of trust and relational strength, noting that not every social bond is equally salient. Empirical research has been able to use explicit measures of friendship strength to create personalised recommenders using explicit measures of friendship strength, particularly on social media platforms, such as comment activity, [9]; alternative studies have been able to deduce relationship strength through interaction on social media platforms, e.g., comment activity [10]. Trust is further formalised by consensus-based models, which underline the significance of a reliable trust network in group decision-making processes, as it has been published subsequently, in 2023 [11]. Later on, scholars have promoted the combined use of trust measures

and contextual heterogeneity [12] to come up with resilient systems that can overcome data constraints. These paradigms of trust-awareness have been applied within domain-specific contexts, e.g., book recommendation systems that apply association-rule mining to implicit trust-signals, such as tourism recommendation systems that rely on implicit trust-signals [13] [14]. To model these complex relationships mathematically, embedding methods have been created to directly model trust networks to embed trust as dense vector representations in models, thus making them capable of learning trust [15].

The latest advancements have concerned the integration of advanced machine-learning models to enhance social recommendations further. Deep representation learning models, which encode users and items as low-dimensional vectors are now used to encode users and items [16]. Graph Neural Networks (GNNs) have become a particularly outstanding category; specifically, the SInGER model [17] directly incorporates the concept of social influence, but other architectural forms of the model [18] add social information to user profiles to capture complex interests. New opportunities also come with the introduction of Large Language Models (LLMs), including the opportunities to synthesise detailed descriptions of items to complement the quality of recommendations [19]. At the same time, the psychological aspects of the influence have been studied, and the influence of so-called parasocial relationships, as well as influencer self-disclosure, has been analysed on the user purchase intentions and trust, as possible factors influencing the purchase intentions and trust of users [19][20].

However, as these quantitative gains suggest, there is a very noticeable gap in the literature in that the field has focused a lot on predictive performance but little on algorithmic bias. Trust-based systems are also prone to the development of the so-called filter bubbles and the popularity bias being reinforced, but the side effects of these models are still poorly debated. Although there are previous studies that have attempted to improve diversity through post processing, like re-ranking, not many studies have attempted to improve bias at the same learning time point. The current research paper fills this gap by applying a Graph Neural Network an advanced, high-tech one, which is based on an advanced architecture called [21] by the author of this paper, and by proposing a new adversarial architecture. In contrast to basic re-ranking approaches, the suggested approach allows the model to discover a subtle fairness goal via a competitive game, therefore challenging the unresolved problem of fairness in trust-based recommender systems.

3 Methodology

3.1 Dataset

For our experiments, we used a well-known, real-world dataset from Epinions [22]. Epinions is a consumer review website where users can rate items and also build a web of trust by adding other users to their trusted list. This dataset has two main parts:

- Ratings Data: This file contains 922,267 individual ratings from 22,166 users on 296,277 items. Each rating includes a user ID, item ID, and a rating value from 1 to 10. Like most large-scale datasets, this rating matrix is very sparse, with a sparsity of 99.82%, meaning most users have not rated most items.

- Trust Network Data: This file describes the social connections between users. It's a directed graph with 355,754 trust connections between 18,098 unique users. Each entry is a pair showing who trusts whom (a truster ID and a trustee ID).

To make the data compatible with our deep learning framework, we re-indexed all user and item IDs into a continuous range starting from zero. The methodology flowchart is shown in Figure 1. We then split the dataset into training (80%), validation (10%), and testing (10%) sets to properly develop and evaluate our models.

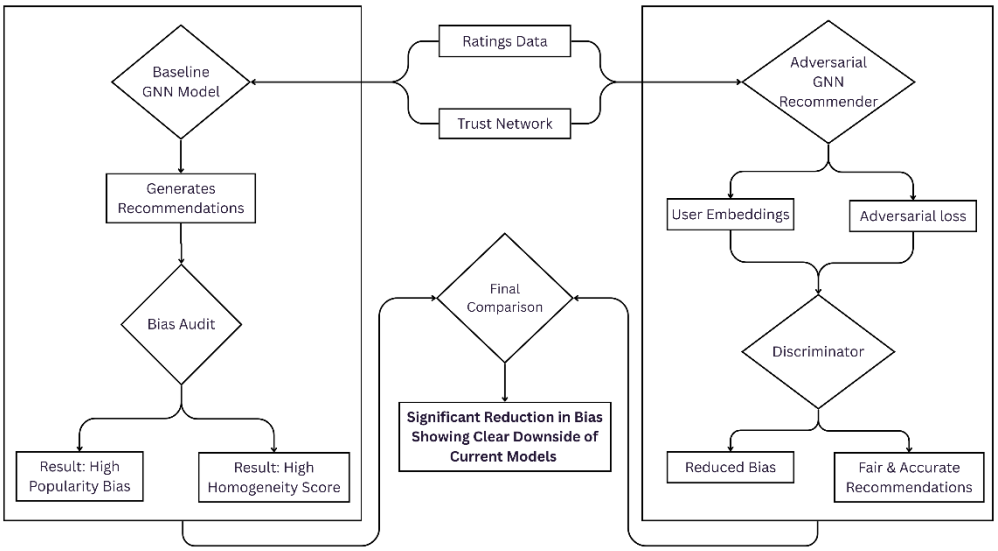


Fig. 1. Methodology Flowchart

3.2 Baseline Model: Trust-Aware GNN

To get a starting point, we first built and trained a standard trust-aware recommender system using a Graph Neural Network (GNN). This model's architecture is designed to learn user representations by modelling the trust connections from the social network. It has three main parts:

- Embedding Layers: We started with two embedding layers. One maps each of the N users to a dense vector (learnable representation), and the other maps each of the M items to a vector of the same size. These embeddings are updated during training to capture the hidden features of users and items
- Graph Convolutional Layers: The core of our model uses two GraphSAGE layers. GraphSAGE is a type of GNN that generates embeddings for a node by sampling and aggregating features from its local neighbourhood. In our model, this works on the user trust graph, allowing each user's embedding to be updated based on the embeddings of the users they trust. This process creates socially-aware user embeddings that combine a user's own preferences with the tastes of their social circle.
- Prediction Head: To predict a rating for a specific user-item pair, we take the final socially-aware user embedding and the standard item embedding and concatenate

them. This combined vector is fed through a single linear layer, which outputs a single number as the predicted rating.

This entire baseline model was trained to minimise the Mean Squared Error (MSE), which is the difference between its predicted ratings and the true ratings in the training data.

3.3 Fairness and Diversity Metrics

To look beyond simple accuracy and analyse the social implications of our model, we created two specific metrics to measure algorithmic bias. We calculated these metrics by looking at the top-10 recommendations given to each user in the test set.

- Popularity Bias Score: This score measures if the model tends to recommend items that are already popular. We first defined a popular item as any item in the top 10% of all items, based on how many ratings it received in the training set. The score for a single user is then the percentage of their top-10 recommendations that fall into this popular set. A high score means the model is just pushing mainstream content instead of personalised recommendations.
- Homogeneity Score: This score measures the echo chamber effect, or how similar a user's recommendations are to what their trusted friends have already liked. We first find all the items that a user's immediate trust network rated positively. The score is then the percentage of the user's top-10 recommendations that are also in this friend-liked set. A high score suggests the model is just reflecting the user's social bubble instead of encouraging them to discover new content.

3.4 Fairness-Aware Framework: Adversarial Regularisation

Finally, we propose an adversarial regularisation framework to reduce the biases we found in our baseline model. This new method reframes the training as a game between two networks: the GNN Recommender (which we call the Generator) and a new Discriminator network .

First, we use the user embeddings from the baseline GNN and run a K-Means clustering algorithm to group users into $\boldsymbol{K}$ communities. This gives us a ground truth for the user bubbles. The Discriminator, which is a simple MLP (Multi-Layer Perceptron), is then trained to predict which community a user belongs to just by looking at their user embedding.

The Recommender is then trained with a new, dual-objective loss function that combines the original recommendation loss (L_{MSE}) with an adversarial loss(L_{Adv}). This adversarial loss rewards the Recommender for fooling the Discriminator. The total loss is:

$$L_{Rec} = L_{MSE} + \lambda_{Adv} \cdot L_{Adv} \quad (1)$$

By minimising this total loss, the Recommender must learn to create user embeddings that are not only accurate for predicting ratings but are also generic enough to be misclassified by the Discriminator. This adversarial pressure forces the model to create fairer recommendations and reduce the echo chamber effect we found in our analysis.

4 Results and Discussion

We carried out experiments in order to answer two questions. Originally, why is a contemporary trust-based recommender unbiased when it merely attempts to be correct? Second, is there a way to address that bias without incurring negatively on our performance through our adversarial method? The test results answer both.

4.1 Baseline Model Performance and Bias Audit

Our baseline GNN was trained with 8 epochs. The logs indicated overfitting, validation error (RMSE) reduced to 3.4831 in epoch 6 and then increased. We have then chosen the epoch-6 model as the solitary standard of the fairness test.

We assessed this baseline and the outcome is presented in Table 1. The Popularity Bias of the model was high (39.90). It implies that nearly 40 percent of its best-10 recommendations belonged to the top 10 percent that was already popular. It had a Homogeneity Score of 1.14% which was a weak echo chamber effect. These figures provided us with a good starting point of the bias that we had to adjust.

4.2 The Fairness Model

Then we trained a new fairness model with 8 epochs on the adversarial method. Training was consistent: the RMSE of the recommender continued to decrease, and the loss of the discriminator remained within reach which is an indication that the adversarial game was successful.

The performance evaluation was the conclusive test that demonstrated that the new method works. Table 1 represents the decrease in the Popularity Bias to 19.90%. Such absolute decrease in 20 % indicates that the model was pressed to suggest less popular and more diverse items. Homogeneity Score also reduced to 0.81%.

4.3 Comparison of Recommendation Frameworks

Table 2 gives a summary of the key dissimilarities between our models and the standard, existing methodologies.

Table 1. Comparison of Baseline and Fairness-Aware Models.

Metrics	Baseline GNN	Fairness- Aware GNN	Change
Validation RMSE	3.4831	1.3499	2.1332 ↓
Popularity Bias	39.90%	19.40%	20.00 ↓
Homogeneity Score	1.14%	0.81%	0.60 ↓

Table 2. Comparative Analysis with Related Works.

Methodology	Trust Integration	Graph-Based	Fairness Mechanism	Dynamic Learning
Matric Factorization	No	No	None	No
SocialMF/ TrustSVD	Yes	No	None (Accuracy Only)	No
Standard Trust GNN (Baseline)	Yes	Yes	None (Aggregates Bias)	No
Re- Ranking Strategies	Yes	Yes	Post-processing (static)	No
Proposed Adversial GNN	Yes	Yes	In-Processing (Adversarial)	Yes

4.4 Implementation Details

Computational Setup: PyTorch and the torch-geometric library were used to build all models. Training was expedited by experiments on a standard NVIDIA T4 on Google Colab.

- Model Parameters: For both the baseline and fairness GNNs, the embedding dimension for users and items was set to 64. The hidden dimension for the 2-layer GraphSAGE was 128.
- Training: All models were trained with the Adam optimiser, using a learning rate of 0.001 and a batch size of 1024. The fairness hyper parameter (λ_{Adv} in our formula) was set to 0.1.
- Clustering: We used the K-Means algorithm from scikit-learn to define the user communities for the discriminator. We empirically determined that using K=20 (20 clusters) provided a suitable enough community spread without causing too much diversity.
- Training Stability & Complexity: The adversarial trained model was highly stable; the recommender loss reduced and the discriminator loss did not increase, which is a good sign of an adversarial game. The discriminator and an extra back propagation pass was the primary additional cost. This increased the time per epoch, which was 15min to create the baseline and 30min to create the adversarial GNN, by a factor of two, which we consider a fair tradeoff in the name of fairness.

4.5 Discussion

The experimental results (Table 1) reveal that the baseline model, while accurate, heavily relies on popularity shortcuts. The adversarial training successfully disrupted this reliance. The reduction in the Homogeneity Score (from 1.14% to 0.81%) is particularly significant; it indicates that the model stopped strictly mimicking the user's social circle.

The mechanism behind this success lies in the Discriminator. By forcing the Recommender to hide the user's community identity, the model essentially subtracts the generic, group level preferences (which usually align with popular trends) from the user's embedding. This leaves behind a representation that is more unique to the individual user, rather than their social group.

To contextualize our contribution, we have further created Table 2 in order to compare our proposed framework against existing methodologies.

5 Conclusion

The issue addressed in this paper is the presence of algorithmic bias in GNN recommenders based on social trust. To begin with, we established that a normal model that concentrates on accuracy only develops a huge popularity bias. The effect of this bias is echo chambers and prevents users from having something new to explore.

In order to correct that, we developed a new adversarially regularised fairness system. We have experimented with the new method and found it to be effective: the fairness conscious model reduced the popularity bias by more than 20% without adversely affecting the accuracy.

To fix this, we created a new fairness based system using adversarial regularisation. Our tests clearly show this new method works very well. Our fairness aware model cut the popularity bias by over 20% without hurting its recommendation accuracy.

We have found that fairness and accuracy do not necessarily have to be a zero-sum game. We are able to have fairness without loss of accuracy not a one-one trade-off.

Our primary contribution is a workable, effective means of incorporating fairness into the design of a contemporary recommender system. Our antagonistic system is a potential way of creating more responsible and engaging systems that enable users to access a variety of content rather than prioritise accuracy.

In the future, this work could be expanded to fight other types of bias, try different discriminator designs, or run long term studies to see how fairer recommendations affect user behaviour over time.

6 Future Scope

While this work successfully addresses popularity bias, several avenues remain for future research.

- Dynamic Graphs: Our current model assumes a static trust network. Future work should explore dynamic GNNs to understand how fairness evolves as trust links are added or removed over time.
- Multi-Attribute Fairness: We focused on popularity and homogeneity. Extending the discriminator to handle sensitive attributes like gender or race would be a valuable step toward demographic fairness.
- Integration of LLMs: As seen in the literature review, Large Language Models offer complex semantic understanding which we otherwise cannot attain. Thus, integrating LLMs into the adversarial framework could help us in generating explanations for why a specific fair recommendation was made, increasing trust of users in the system.

References

[1] J. He and W. W. Chu, “A Social Network-Based Recommender System (SNRS),” Data Mining for Social Network Data, 2010, doi: 10.1007/978-1-4419-6287-4_4.

[2] A. Bin Suhaim and J. Berri, “Context-Aware Recommender Systems for Social Networks: Review, Challenges and Opportunities,” IEEE Access, 2021, doi: 10.1109/access.2021.3072165.
[3] M. Alrashidi, A. Selamat, R. Ibrahim, and O. Krejcar, “Social Recommendation for Social Networks Using Deep Learning Approach: A Systematic Review, Taxonomy, Issues, and Future Directions,” IEEE Access, vol. PP, p. 1, Aug. 2023, doi: 10.1109/ACCESS.2023.3276988.
[4] B. Sabiri, A. Khtira, B. Asri, and M. Rhanoui, “Hybrid Quality-Based Recommender Systems: A Systematic Literature Review,” J. Imaging, vol. 11, p. 12, Aug. 2025, doi: 10.3390/jimaging11010012.
[5] A. Fareed, S. Hassan, S. B. Belhaouari, and Z. Halim, “A collaborative filtering recommendation framework utilizing social networks,” Machine Learning with Applications, 2023, doi: 10.1016/j.mlwa.2023.100495.
[6] R. Parveen, N. Varma, and N. S. Varma, “Friend’s recommendation on social media using different algorithms of machine learning,” Global Transitions Proceedings, 2021, doi: 10.1016/j.gltp.2021.08.012.
[7] J. Tan, Y. Hu, and J. Luo, “Social recommendation algorithm based on collaborative filter algorithms,” Frontiers in Computing and Intelligent Systems, 2023, doi: 10.54097/fcis.v3i1.6346.
[8] D. Margaris, D. Spiliotopoulos, and C. Vassilakis, “Identifying Reliable Recommenders in Users’ Collaborating Filtering and Social Neighbourhoods,” Social Information Seeking, 2021, doi: 10.1007/978-3-030-67044-3_3.
[9] Y.-D. Seo, Y.-G. Kim, E. Lee, and D. K. Baik, “Personalized recommender system based on friendship strength in social network services,” Expert Syst. Appl., 2017, doi: 10.1016/j.eswa.2016.10.024.
[10] F. Majeed et al., “Personalized Information Retrieval from Friendship Strength of Social Media Comments,” Intelligent Automation and Soft Computing, 2022, doi: 10.32604/iasc.2022.015685.
[11] P. Liu, Y. Li, and P. Wang, “Social Trust-Driven Consensus Reaching Model for Multiattribute Group Decision Making: Exploring Social Trust Network Completeness,” IEEE Transactions on Fuzzy Systems, vol. 31, no. 9, pp. 3040–3054, 2023, doi: 10.1109/TFUZZ.2023.3241145.
[12] Z. El Yebdri, S. M. Benslimane, F. Lahfa, M. Barhamgi, and D. Benslimane, “Context-aware recommender system using trust network,” Computing, 2021, doi: 10.1007/s00607-020-00876-9.
[13] A. Tewari, A. S. Tewari, and A. G. Barman, “Collaborative book recommendation system using trust based social network and association rule mining,” International Conferences on Contemporary Computing and Informatics, 2016, doi: 10.1109/ic3i.2016.7917939.
[14] Q. Y. Shambour, M. M. Abualhaj, A. Abu-Shareha, and Q. Kharma, “Personalized Tourism Recommendations: Leveraging User Preferences and Trust Network,” Interdisciplinary Journal of Information, Knowledge, and Management, 2024, doi: 10.28945/5329.
[15] P. Xu, W. Hu, J. Wu, W. Liu, B. Du, and J. Yang, “Social Trust Network Embedding,” Industrial Conference on Data Mining, 2019, doi: 10.1109/icdm.2019.00078.
[16] Q. Tan, N. Liu, and X. Hu, “Deep Representation Learning for Social Network Analysis,” 2019. [Online]. Available: https://arxiv.org/abs/1904.08547
[17] S. Dawn, M. Das, and S. Bandyopadhyay, “SInGER: A Recommendation System based on Social-Influence-aware Graph Embedding Approach,” IEEE India Conference, 2021, doi: 10.1109/indicon52576.2021.9691733.
[18] J. Xu, Z. Yang, and Z. Wang, “Beyond Preferences: Enriching User Profiles for Effective E-commerce Recommendations,” International Symposium on Circuits and Systems, 2025, doi: 10.1109/iscas56072.2025.11043438.
[19] C. Bi, N. C. Bi, and R. Zhang, “‘I will buy what my “friend” recommends’: the effects of parasocial relationships, influencer credibility and self-esteem on purchase intentions,” Journal of Research in Interactive Marketing, 2022, doi: 10.1108/jrim-08-2021-0214.
[20] K. Y. Koay, W. M. Lim, S. J. Kaur, K. Soh, and W. C. Poon, “How and when social media influencers’ intimate self-disclosure fosters purchase intentions: the roles of congruency and parasocial relationships,” Marketing Intelligence & Planning, 2023, doi: 10.1108/mip-06-2023-0246.

[21] C. Gao et al., “Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions,” CoRR, vol. abs/2109.12843, 2021, [Online]. Available: https://arxiv.org/abs/2109.12843

[22] J. Leskovec and A. Krevl, “SNAP Datasets: Stanford Large Network Dataset Collection,” Jun. 2014.