# A Comprehensive Evaluation Framework for Multi-Level Bias Analysis in Graph-Based Personalization Systems

Hrushikesh Ahire[1], Gavin Rony Correia[1], Pinky Sherwani[1], Het Darshan Mehta[1,2], Marco Polignano[3,1], Giovanni Semeraro[3] and Ernesto William De Luca[1,2]

[1]*Otto-von-Guericke University, Universitatplatz 2, 39106 Magdeburg, Germany*
[2]*Leibniz Institute for Educational Media | Georg Eckert Institute, Freisestr. 1, 38118 Braunschweig, Germany*
[3]*Università degli studi di Bari Aldo Moro, Piazza Umberto I, 70121 Bari, Italy*

**Abstract**

Social media platforms rely on algorithmic curation to rank and recommend content, thereby shaping user exposure and public discourse. While these systems improve personalization, they can also introduce algorithmic bias, reinforce echo chambers, and intensify political polarization. This paper presents a comprehensive evaluation framework for analyzing multi-level bias in graph-based personalization systems. We model social media interactions as graphs and use Graph Neural Networks (GNNs) to learn user representations. Rather than proposing a new model, we compare a vanilla GNN with two fairness-aware methods, FairGNN and FairVGNN, within a unified evaluation pipeline. We further incorporate a controlled recommender simulation to assess how exposure affects fairness and polarization, and an explainability layer that combines GNNExplainer with a Large Language Model to generate human-readable audit reports. Results show that fairness-aware models can reduce outcome-level bias, but may still retain representation-level clustering and remain sensitive to exposure effects, highlighting the importance of pipeline-level fairness evaluation.

**Keywords**

Explainable User Models, Fairness-Aware Graph Learning, Transparent Personalization, Graph Neural Networks, Social Media Bias, Explainable Recommendation

## 1. Introduction

Social media platforms rely heavily on algorithmic curation to personalize content and user interactions. These systems determine which information is shown to users, thereby shaping visibility, engagement, and ultimately public discourse [1]. While personalization improves user experience, it also introduces critical challenges related to algorithmic bias, echo chambers, and polarization. Algorithmic bias can lead to systematically unequal outcomes across user groups, often reflecting or amplifying structural patterns present in the underlying data [2]. In social networks, these effects are closely linked to homophily, where users preferentially connect with others who share similar attributes or beliefs. This process contributes to the formation of echo chambers, in which users are primarily exposed to homogeneous viewpoints, limiting exposure to diverse perspectives [3, 4, 5]. Prior work has shown that such dynamics can fragment network structures and reinforce polarization through the interaction of user behavior and algorithmic ranking mechanisms [6, 7]. Beyond fairness concerns, a key limitation of modern recommender systems is their lack of transparency. Many models operate as black boxes, making it difficult for users and stakeholders to understand how recommendations are generated and how bias emerges. Explainable Artificial Intelligence (XAI) aims to address this issue by providing interpretable insights into model behavior, thereby improving trust and accountability [8, 9]. In personalized systems, explainability is particularly important, as it enables users to understand not only what is recommended, but also why it is recommended and how potential biases influence these outcomes. Despite substantial

*Joint Proceedings of the ACM UMAP Workshops 2026, UMAP 2026, 8–11, 2026, Gothenburg, Sweden*
hrushikesh.ahire@st.ovgu.de (H. Ahire); gavin.correia@st.ovgu.de (G. R. Correia); pinky.sherwani@st.ovgu.de (P. Sherwani); het.mehta@ovgu.de (H. D. Mehta); marco.polignano@uniba.it (M. Polignano); giovanni.semeraro@uniba.it (G. Semeraro); ernesto.deluca@ovgu.de (E. W. De Luca)
0009-0008-1924-823X (P. Sherwani); 0009-0000-8372-3877 (H. D. Mehta); 0000-0002-3939-0136 (M. Polignano); 0000-0001-6883-1853 (G. Semeraro); 0000-0003-3621-4118 (E. W. De Luca)

progress in fairness-aware learning and explainability methods, these aspects are typically studied in isolation. Existing work often evaluates fairness at the prediction level, or explainability at the local model level, without considering how bias emerges and propagates across the full recommendation pipeline. In particular, there is limited understanding of how bias evolves across (i) the original graph structure, (ii) learned representations, and (iii) post-exposure recommendation dynamics. Furthermore, explainability methods such as GNNExplainer are rarely integrated into a unified framework that connects structural bias, model behavior, and user-facing explanations.

To address these gaps, we propose a comprehensive evaluation framework for analyzing multi-level bias in graph-based personalization systems. Our approach models social media interactions as graphs and applies both standard and fairness-aware Graph Neural Networks (GNNs) [10] to learn user representations. We then simulate a recommendation process to study how exposure mechanisms influence bias amplification. Finally, we incorporate an explainability layer that combines GNNExplainer with a Large Language Model (LLM) to generate human-readable audit reports. The key contribution of this work is to design of a unified evaluation framework that enables systematic analysis of bias across multiple stages of graph-based personalization pipelines. More specifically, the framework supports analysis at three complementary levels: (i) structural bias in the original network, (ii) representation-level bias in learned embeddings, and (iii) outcome-level bias in predictions and recommendations. By explicitly linking these levels, the framework provides a holistic perspective on how bias emerges and propagates in personalized systems, aligning with the goals of explainability and responsible personalization. Based on this framework, we address the following research questions:

- **RQ1:** How can structural bias and polarization in social networks be quantitatively characterized prior to any learning or recommendation process?
- **RQ2:** How do fairness-aware GNN models influence bias across prediction and representation levels?
- **RQ3:** How does similarity-based recommendation affect the amplification or mitigation of bias and polarization under controlled exposure?
- **RQ4:** To what extent can graph-based explanations, enhanced with LLM-generated summaries, support the interpretability of bias and recommendation behavior in personalized systems as a qualitative audit layer?

The proposed framework is intended as a diagnostic audit tool for researchers, system designers, and platform stakeholders. Its practical value lies in identifying where bias appears in a graph-based personalization pipeline: in the original graph structure, in learned embeddings, or after recommendation exposure. Such diagnosis can inform whether mitigation should target data collection, representation learning, ranking, or explanation/reporting. We therefore treat fairness as a pipeline-level property rather than a single output metric.

## 2. Related Work

Understanding bias and transparency in graph-based personalization systems requires insights from multiple research areas, including social network analysis, fairness in machine learning, recommender systems, and explainable AI. Social networks are known to exhibit structural properties such as homophily, where users preferentially connect to similar others, leading to clustered and often polarized communities [11, 7]. These structures can give rise to echo chambers, where users are repeatedly exposed to similar viewpoints, reinforcing existing beliefs and limiting diversity of information [6, 12]. Such structural biases form the underlying substrate on which learning algorithms operate, suggesting that bias in personalized systems cannot be fully understood without analyzing the original graph structure. Graph Neural Networks (GNNs) have become a dominant approach for modeling such data, as they learn representations by aggregating information from node features and graph structure [13, 14]. However, this same mechanism makes GNNs susceptible to encoding and amplifying structural biases present in the graph [15, 16]. Even when sensitive attributes are not explicitly used, they can

be implicitly encoded in the learned representations. This highlights the need to evaluate not only predictive outcomes, but also representation-level bias within learned embedding spaces.

To address these issues, fairness-aware GNN methods such as FairGNN and related approaches introduce mechanisms like adversarial debiasing and feature masking to reduce sensitive attribute leakage [17, 18]. While these methods can improve prediction-level fairness, recent studies show that they may not fully eliminate representation-level bias or structural segregation. This limitation motivates the need for evaluation frameworks that consider fairness across multiple levels rather than focusing solely on prediction metrics. Beyond representation learning, recommender systems play a crucial role in shaping user exposure. Prior work has shown that recommendation algorithms can amplify existing preferences and biases by repeatedly exposing users to similar content, thereby reinforcing homophily and polarization [19]. This suggests that bias cannot be fully assessed at the model level alone, but must also be evaluated under exposure dynamics introduced by recommendation mechanisms. Finally, explainability has emerged as a key requirement for transparent and trustworthy AI systems. Techniques such as GNNExplainer provide local explanations by identifying subgraphs and features that influence predictions [20]. More recently, large language models (LLMs) have been used to translate such technical explanations into human-readable narratives, enabling broader accessibility and interpretability [21]. However, existing approaches typically focus on explanation generation in isolation and do not explicitly connect explanations to fairness analysis or exposure-driven bias.

In contrast to prior work, our approach integrates these perspectives into a unified evaluation framework that combines graph-based learning, recommender simulation, and explainability. This enables a systematic analysis of how bias emerges across structural, representation, and exposure levels, and how these effects can be communicated through interpretable and actionable insights.

## 3. Framework Overview

The proposed pipeline integrates three key components: graph representation learning, recommendation simulation, and explainability-driven auditing, rather than introducing a new graph learning model. This design enables a systematic evaluation of bias at three levels: before learning (structural bias), after representation learning (embedding-level bias), and after simulated exposure (outcome-level bias). All learning components used in this framework are well known approaches in literature, intentionally treated as interchangeable modules to support a controlled and reproducible evaluation.

Figure 1 illustrates the framework pipeline. It begins with attributed social network graphs, where users are represented as nodes and social relations as edges. These graphs are processed by graph learning models to obtain node embeddings and prediction outputs. The learned representations are then used in a controlled recommendation simulation to study how exposure mechanisms influence bias amplification. Finally, an explainability layer combines GNNExplainer with LLM-based report generation to translate model behavior into human-readable audit summaries. This staged design allows us to connect structural properties of the graph, learned representations, and exposure dynamics within a single evaluation framework. In this sense, the proposed approach provides a comprehensive pipeline for multi-level bias analysis in graph-based personalized systems.

### Datasets and Preprocessing

We conduct our analysis on the Pokec social network benchmarks [22], specifically Pokec-z and Pokec-n, which are widely used in fairness-aware graph learning research. Pokec is a Slovakian online social network where users create profiles and establish friendship links. In both datasets, nodes correspond to users and edges correspond to social relations, resulting in attributed graphs that combine structural information, profile features, target labels, and sensitive attributes.

In our experimental setup, we consider two complementary fairness scenarios. For Pokec-z, the sensitive attribute corresponds to *region*, while for Pokec-n it corresponds to *gender*. The prediction targets are defined as binary attributes derived from user profiles: *working field* for Pokec-z and *age* for Pokec-n. This design enables evaluation across different types of sensitive attributes and target variables,

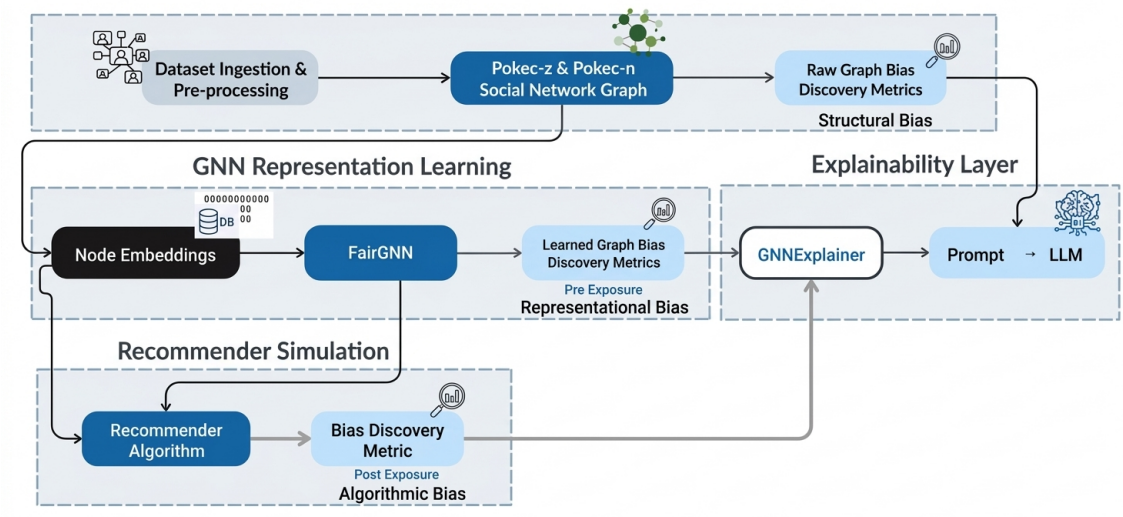



**Figure 1:** Overview of the proposed framework for multi-level bias analysis in graph-based personalization systems.

| Dataset | # Nodes | # Edges | Sensitive Attribute ($S$) | Target Attribute ($Y$) |
|---|---|---|---|---|
| **Pokec-z** | 67,796 | 1,303,712 | Region<br>$S = 0$: Outside the capital region [65.61%]<br>$S = 1$: Within the capital region [34.39%] | Working Field<br>$Y = 0$: Not working in the field [68%]<br>$Y = 1$: Working in the field [32%] |
| **Pokec-n** | 66,569 | 1,100,663 | Gender<br>$S = 0$: Female [51.28%]<br>$S = 1$: Male [48.72%] | Age<br>$Y = 0$: $\leq 25$ years old [58%]<br>$Y = 1$: $> 25$ years old [42%] |

**Table 1**
Dataset statistics and fairness-relevant attribute definitions.

allowing the framework to generalize across heterogeneous fairness settings. The class distributions for the target variables are reported approximately, as they may vary slightly depending on preprocessing and filtering steps. These datasets are particularly suitable for our study as they provide large-scale graph structure together with explicit sensitive attributes, enabling controlled evaluation of fairness and bias under different structural conditions. Notably, Pokec-z exhibits a strong imbalance in the sensitive attribute distribution and is highly homophilous, while Pokec-n presents a near-balanced sensitive attribute distribution and more mixed connectivity patterns. This distinction allows us to analyze how underlying graph structure influences bias propagation across the pipeline. Table 1 summarizes the dataset statistics and attribute definitions.Before training, node features are normalized to ensure numerical stability, while sensitive attributes are preserved explicitly for fairness-aware evaluation. The graph is represented as a sparse adjacency matrix for scalability. Train, validation, and test splits are created at the node level, ensuring consistent alignment between features, labels, and sensitive attributes across all models.

In Pokec-z, Region is treated as a sensitive attribute because geographic location can act as a proxy for socioeconomic status, access to opportunities, and urban/rural visibility. In people-to-people recommendation, overexposure to users from the same region may reinforce geographic segregation and reduce cross-regional discoverability. The target attribute, Working Field, is used as the downstream prediction task through which we study whether regional membership affects model behavior. In Pokec-n, Gender is treated as a sensitive attribute following standard fairness evaluation practice in graph learning. The target attribute, Age, is used as a binary prediction label. The goal is not to claim that all uses of these attributes are harmful, but to evaluate whether graph-based personalization produces systematically different predictions or exposure patterns across protected or socially meaningful user groups.

### Models

To analyze how bias is reflected in learned representations, we compare one standard graph model and two fairness-aware models: a vanilla GNN [23], FairGNN [24], and FairVGNN [25]. All models share a Graph Convolutional Network (GCN) backbone to ensure comparability across conditions. The vanilla GNN serves as a baseline, capturing how standard graph learning propagates structural properties of the network into embeddings and predictions. In contrast, FairGNN introduces adversarial debiasing to reduce sensitive attribute leakage in learned representations [17, 26]. A discriminator attempts to predict the sensitive attribute, while the encoder is trained to preserve task-relevant information while suppressing sensitive signals. FairVGNN extends this approach by incorporating variational feature masking [26]. It identifies features correlated with the sensitive attribute and learns soft masks to suppress them during training. This results in representations that aim to balance predictive performance and fairness through selective information filtering. Across all models, hyperparameters are selected on the validation split via grid search, and results are reported over multiple runs. Importantly, these models are not proposed as novel contributions, but are used as controlled components within the broader evaluation framework to study how bias evolves across different stages of the pipeline.

## 4. Recommender Simulation

To move beyond prediction-level analysis, we use a recommendation simulation layer that models user exposure based on learned representations. This component is central to the framework because it allows us to examine how bias is not only encoded by graph learning models, but also propagated or amplified once these models are placed in a recommendation pipeline. In this sense, the recommender is not treated as an optimization target, but as a controlled evaluation mechanism for studying exposure effects. The simulation uses node embeddings as latent user representations. For each user, pairwise similarity scores are computed against all other users using cosine similarity. We adopt cosine similarity because it captures directional proximity in the embedding space and is widely used in representation-based retrieval and nearest-neighbor recommendation. Users are ranked according to these similarity scores, and the top-$k$ most similar users are selected as recommendations. This setup approximates common social recommendation scenarios such as ”people you may know” or similarity-driven user exposure, while remaining simple enough to support controlled comparison across graph learning conditions. A key design choice of this component is that the ranking strategy is intentionally fixed. The purpose of the recommendation simulation is not to maximize recommendation quality, but to isolate how differences in learned embeddings translate into downstream exposure patterns. By keeping the recommendation mechanism constant across models, the framework can attribute differences in post-exposure behavior to the representations themselves rather than to changes in recommender architecture. In addition to similarity-based ranking, the simulation retains prediction outputs from the graph models, including predicted labels and confidence scores. These outputs are used to analyze how recommendation interacts with model confidence, group membership, and fairness-related behavior. Sensitive attributes are never used for ranking, but are retained for auditing. This allows direct comparison between pre-exposure and post-exposure states and makes it possible to evaluate whether fairness-aware representation learning remains effective once a model is embedded in a similarity-based exposure process. Overall, the recommender simulation serves as the exposure layer of the evaluation framework. It links representation learning to user-facing recommendation behavior and provides the basis for studying how graph-based personalization systems may unintentionally reinforce homophily, echo chambers, and polarization.

## 5. Explainability and LLM-Based Auditing

To complement quantitative analysis, the framework includes an explainability and auditing layer that translates model behavior into interpretable insights. This component is central to the framework

because it connects graph-based predictions and recommendation behavior with human-readable explanations. Rather than proposing a new explainability method, it integrates existing explanation techniques into the evaluation pipeline in order to support transparency, bias auditing, and stakeholder-facing interpretation. At the local level, we apply GNNExplainer to identify the subgraph structure and feature dimensions that are most influential for a given prediction [20]. This produces two complementary forms of explanation. First, structural attribution highlights the edges in the local neighborhood that contribute most strongly to the prediction or recommendation. Second, feature attribution identifies the node attributes that have the greatest influence on the model output. Together, these explanations make it possible to determine whether recommendations are primarily driven by direct graph connectivity, indirect structural similarity, or specific profile-level features.

Beyond local explanations, the framework performs fairness-oriented auditing at the system level. In particular, we compute the homophily ratio of recommendation lists, defined as the proportion of recommended users who share the same sensitive attribute as the target user. High homophily ratios indicate the potential reinforcement of same-group exposure and possible filter-bubble behavior. We additionally compare prediction score distributions across sensitive groups in order to identify systematic disparities in model confidence and group-conditioned outcomes. To make these signals accessible to non-technical stakeholders, we further incorporate a Large Language Model (LLM) for natural-language reporting. The LLM receives structured inputs derived from explanation and auditing metrics, including structural explanation patterns, feature attribution scores, recommendation confidence, and homophily-based indicators. Based on these inputs, it generates concise summaries that describe the likely drivers of a recommendation, the degree of exposure diversity, and the potential risk of bias or echo-chamber reinforcement. This integration enables the framework to bridge the gap between technical model analysis and stakeholder-facing transparency. By combining local explanations, fairness auditing, and natural-language summaries, the explainability layer extends beyond model inspection and functions as an interpretable reporting mechanism for bias analysis in graph-based personalized systems.

## 6. Experimental Protocol and Results

### Experimental Setup

The empirical study is designed to evaluate bias at three stages of a graph-based personalization pipeline: before representation learning, after representation learning, and after simulated recommendation exposure. Experiments are conducted on the two Pokec benchmarks introduced earlier, namely Pokec-z and Pokec-n, using the same preprocessing pipeline and graph construction settings across all conditions. We use three graph learning models within the framework: a vanilla GNN, FairGNN, and FairVGNN. For all three methods, we adopt a GCN backbone in order to keep the encoder architecture fixed and isolate the effect of fairness-aware learning. Hyperparameters are selected on the validation set through grid search, and final scores are averaged over three runs. These models are not introduced as novel contributions, but are treated as analytical components within the framework. This controlled setup allows us to study how different representation learning strategies affect structural, representation-level, and outcome-level bias.

The recommendation stage is simulated on top of the learned node embeddings. For each user, we compute pairwise similarity scores between the target embedding and all other user embeddings using cosine similarity. Cosine similarity is appropriate in this setting because it captures directional proximity in latent space and is widely used in representation-based retrieval and nearest-neighbor recommendation. Users are then ranked by similarity, and the top-$k$ most similar users are selected as recommendations. In the final implementation, we set $k = 10$, providing a controlled approximation of a limited exposure set in social recommendation scenarios. Self-recommendations and already connected neighbors are excluded from the recommendation candidate set. The recommendation stage is used as a controlled exposure mechanism rather than a fully optimized recommender system. Its purpose is to compare pre-exposure and post-exposure fairness and polarization under a fixed ranking rule, thereby isolating how learned representations translate into downstream exposure patterns. For the

explainability layer, we apply GNNExplainer to selected model outputs in order to identify the most influential local subgraph structure and feature dimensions behind individual recommendations. These explanation signals are then summarized through a Large Language Model to produce short audit-style reports. For LLM-based report generation, we use `llama3.2` in a zero-shot setting. The prompt is structured and includes the model identifier together with fairness and exposure indicators extracted from the recommendation output, including pre-exposure and post-exposure sensitive edge ratios, bias amplification, exposure disparity, and recommendation accuracy. The model is instructed to generate a concise technical audit report assessing whether the recommendation process amplified or mitigated bias. The full prompt template is provided in Appendix B. Generation is performed with temperature = 0.7, top-$p = 0.9$, top-$k = 40$, and a maximum output length of 200 tokens. The train/validation/test split ratio is 70/15/15, and all reported results are averaged across three runs. To assess whether observed differences between models are robust across random seeds, we additionally conduct paired Wilcoxon signed-rank tests on matched seed-level results. Values are reported as median [min, max] over three random seeds. Paired Wilcoxon signed-rank tests were conducted across matched seeds for all model pairs. For each dataset and metric, we compare model pairs using the same three seeds, i.e., GNN vs. FairGNN, GNN vs. FairVGNN, and FairGNN vs. FairVGNN. We use a significance threshold of $\alpha = 0.05$. These paired runs tests are interpreted as exploratory robustness checks and a definitive evidence of statistical significance.

## Evaluation Metrics

To evaluate the framework, we use metric families that correspond to the three levels of analysis considered in this paper: predictive fairness, structural and representation-level bias, and opinion polarization. Unless otherwise stated, these metrics are computed both before and after the recommendation stage in order to assess how exposure changes the bias profile of each model. For downstream node classification, we report group-wise accuracy, AUC-ROC, and F1-score. To evaluate predictive fairness, we use Statistical Parity ($\Delta SP$), Equal Opportunity ($\Delta EO$), and AUC Parity ($\Delta AUC$) [27]. Overall accuracy is computed as the average of the group-wise accuracies:

$$Acc = \frac{ACC_{sen0} + ACC_{sen1}}{2} \tag{1}$$

Statistical Parity measures whether different sensitive groups receive positive predictions at similar rates:

$$\Delta SP = |P(\hat{y}_u = 1 \mid s = 0) - P(\hat{y}_u = 1 \mid s = 1)| \tag{2}$$

Equality of Opportunity measures whether users with positive ground-truth labels are treated similarly across sensitive groups:

$$\Delta EO = |P(\hat{y}_u = 1 \mid y_u = 1, s = 0) - P(\hat{y}_u = 1 \mid y_u = 1, s = 1)| \tag{3}$$

AUC Parity measures disparities in ranking quality across sensitive groups:

$$\Delta AUC = |AUC_{s=0} - AUC_{s=1}| \tag{4}$$

To quantify structural bias and echo chambers, we analyze the original graph and, later, a learned $k$-nearest-neighbor graph constructed from the embeddings. We report intra-group edges, inter-group edges, the intra-group edge proportion, the E–I index, and attribute assortativity. The E–I index is defined as:

$$\frac{\text{inter-group edges} - \text{intra-group edges}}{\text{total edges}} \tag{5}$$

Lower E-I values indicate stronger within-group connectivity and therefore stronger structural segregation. To characterize representation-induced and post-exposure polarization, we additionally

report probability variance, extreme probability mass, group mean gap, false positive rate gap, decision polarization, and Jensen–Shannon divergence between group-conditioned score distributions [17]. The false positive rate gap is computed as:

$$\Delta FPR = |P(\hat{y}_u = 1 \mid y_u = 0, s_u = 0) - P(\hat{y}_u = 1 \mid y_u = 0, s_u = 1)| \tag{6}$$

and the Jensen–Shannon divergence is computed as:

$$JS(P_0 \parallel P_1) = \frac{1}{2}KL(P_0 \parallel M) + \frac{1}{2}KL(P_1 \parallel M), \quad \text{where } M = \frac{1}{2}(P_0 + P_1) \tag{7}$$

Together, these metrics allow us to assess whether fairness-aware learning improves only outcome-level fairness or also affects representation quality, exposure diversity, and polarization after recommendation.

## Pre-Exposure Analysis

We begin by analyzing the raw graph structure and the learned representations before recommendation exposure. The two datasets exhibit markedly different structural properties. The raw Pokec-n graph shows moderate mixing between sensitive groups, with negative sensitive-attribute assortativity, indicating relatively weak homophily and limited structural echo chambers. In contrast, Pokec-z exhibits very strong within-group concentration, with more than 95% of edges occurring within the same sensitive group and very high assortativity. This indicates that Pokec-z is already highly polarized before any model is applied. Table 2 reports predictive performance and fairness metrics across the three graph learning models. On Pokec-z, FairGNN achieves the highest accuracy, while FairVGNN provides the strongest AUC-ROC and F1-score. In terms of fairness, FairGNN yields the lowest $\Delta SP$ and $\Delta EO$, suggesting the strongest outcome-level fairness among the three methods in this setting. On Pokec-n, FairGNN again achieves the best accuracy and the lowest $\Delta SP$, whereas FairVGNN provides the strongest AUC-ROC and F1-score. Overall, these results indicate that fairness-aware learning affects predictive and fairness-related behavior differently across datasets and metrics. Two observations warrant closer examination before proceeding.

First, the absolute values of $\Delta_{SP}$ and $\Delta_{EO}$ are consistently high across all three methods on both datasets, frequently exceeding 0.50 even for fairness-aware models. This is not indicative of model failure but reflects a structural ceiling imposed by the data itself. On Pokec-z, more than 95% of edges connect nodes within the same sensitive group (Table 3 and 4, raw graph $p_{intra} = 0.9506$), and the target label (*Working Field*) is substantially correlated with *Region* at the population level. When GCN message passing aggregates features across this highly homophilous graph, it propagates group-homogeneous signals throughout the network from the very first layer, making the predicted label rates for the two groups naturally divergent regardless of the training objective. Even adversarial debiasing cannot fully decouple predictions from a sensitive attribute that is so thoroughly encoded in the graph topology — a finding that itself motivates the need for structural-level pre-exposure analysis as a necessary baseline before interpreting any outcome-level fairness metric.

Second, FairVGNN on Pokec-n produces higher $\Delta_{SP}$ (0.4994) and $\Delta_{EO}$ (0.8338) than the vanilla GNN (0.5032 and 0.7158, respectively), an apparent paradox for a method explicitly designed to improve fairness. This inversion reflects a known tension in feature-masking approaches. FairVGNN identifies features correlated with the sensitive attribute (*Gender* in Pokec-n) and learns soft masks to suppress them. However, some of those features may have been functioning as compensatory signals — features that inadvertently equalized predicted outcomes across groups by partially offsetting structural disadvantages for the minority group. Suppressing them removes a fairness-beneficial side effect, widening the prediction gap even as the method successfully reduces sensitive-attribute leakage in the embedding space. Crucially, Table 4 confirms that FairVGNN achieves lower extreme mass (0.0698 vs. 0.1087) and lower Jensen–Shannon divergence (0.0022 vs. 0.0083) on Pokec-n compared to the vanilla GNN, meaning that its debiasing is genuine at the representation level while paradoxically increasing prediction-level disparity. This dissociation between representation-level and prediction-level fairness is

precisely the type of cross-level inconsistency that the proposed framework is designed to surface — and that single-level evaluation would fail to detect.

| Dataset | Method | Accuracy (↑) | AUC-ROC (↑) | F1-Score (↑) | ΔSP (↓) | ΔEO (↓) |
|---|---|---|---|---|---|---|
| Pokec-z | GNN | 0.6699 [0.640, 0.673] | 0.6726 [0.660, 0.675] | 0.6699 [0.658, 0.673] | 0.6879 [0.676, 0.688] | 0.6855 [0.673, 0.687] |
| | FairGNN | **0.6781** [0.667, 0.678] | 0.6810 [0.673, 0.681] | 0.6781 [0.667, 0.678] | 0.6836 [0.660, 0.684] | **0.6810** [0.652, 0.681] |
| | FairVGNN | 0.6431 [0.611, 0.691] | **0.7215** [0.707, 0.752] | **0.6900** [0.684, 0.720] | **0.6438** [0.517, 0.714] | 0.7035 [0.692, 0.724] |
| Pokec-n | GNN | 0.6930 [0.673, 0.693] | 0.5068 [0.507, 0.512] | 0.6930 [0.693, 0.693] | 0.5032 [0.503, 0.508] | **0.7158** [0.715, 0.716] |
| | FairGNN | 0.6987 [0.698, 0.699] | 0.5002 [0.500, 0.504] | 0.6987 [0.698, 0.699] | 0.500 [0.500, 0.503] | 0.7211 [0.720, 0.721] |
| | FairVGNN | **0.7046** [0.695, 0.715] | **0.5753** [0.568, 0.588] | **0.8248** [0.818, 0.838] | **0.4994** [0.365, 0.565] | 0.8338 [0.834, 0.834] |

**Table 2**
Predictive performance and fairness metrics across datasets and models. Best values are highlighted in **Bold**, (↑) indicates that higher values are better, (↓) indicates that lower values are better. Values are reported as median [min, max] over three random seeds with significance at $\alpha = 0.05$

Table 3 and 4 provides a detailed characterization of structural bias and representation-level polarization before recommendation exposure. The reported metrics capture complementary aspects of group segregation and distributional imbalance in both the raw graph and the learned embedding space. In particular, the intra-group edge proportion ($p_{intra}$), E–I index, and assortativity quantify the extent of homophily and structural clustering, while probability variance, extreme mass, group mean gap, and Jensen–Shannon divergence capture how prediction confidence and decision distributions differ across sensitive groups. Probability variance and extreme probability mass capture whether model confidence becomes concentrated. Group mean gap and Jensen–Shannon divergence measure separation between group-conditioned score distributions. False positive rate gap captures group-level asymmetry in final model decisions.These metrics are critical for our evaluation framework because they allow us to distinguish between two fundamentally different types of bias: structural bias inherited from the graph topology and representation-level bias introduced or amplified during learning. This distinction directly addresses RQ1 and RQ2, as it enables us to assess whether fairness-aware models reduce only outcome-level disparities or also mitigate deeper structural and embedding-level segregation. On Pokec-n, both FairGNN and FairVGNN slightly increase homophily in the learned graph relative to the raw structure, as indicated by higher $p_{intra}$ and assortativity values, although the overall level of polarization remains relatively low. In contrast, Pokec-z exhibits strong and persistent structural segregation across all models. While representation learning reduces assortativity compared to the raw graph, the learned embeddings remain highly clustered, indicating that fairness-aware training does not fully overcome the strong homophily inherent in the original network.

Furthermore, polarization metrics such as extreme probability mass and Jensen–Shannon divergence reveal that FairVGNN tends to produce more concentrated prediction distributions in Pokec-z, suggesting stronger confidence polarization despite competitive predictive performance. This highlights a key insight of our framework: improvements in predictive fairness metrics do not necessarily imply reductions in representation-level bias or polarization.

Table 3 and 4 reveals a further and more counterintuitive finding: on Pokec-z, the vanilla GNN produces the most structurally mixed embedding space of all three models, reducing the intra-group edge proportion in the learned k-NN graph from 0.9506 (raw graph) to 0.5580, while FairGNN and FairVGNN retain far higher clustering (0.8262 and 0.8860, respectively). This inversion, where the non-fairness-aware baseline is more structurally diverse at the representation level than the fairness-aware methods, arises from a fundamental difference in training objectives. The vanilla GNN optimizes purely for the classification target (Working Field), and since this label does not map perfectly onto Region at the individual level in Pokec-z, the gradient signal pushes node embeddings toward task-relevant clusters that cross group boundaries. FairGNN and FairVGNN, by contrast, must explicitly model the sensitive attribute in order to suppress it: FairGNN's adversarial discriminator and FairVGNN's feature masking both require the encoder to maintain awareness of group structure during training, which

reinforces the geometric organization of the embedding space around sensitive group membership even while reducing direct sensitive-attribute predictability. In other words, these methods suppress what group membership predicts about outcomes without necessarily dissolving the geometric proximity structure that reflects group membership in the first place. This finding directly validates the need for representation-level evaluation: outcome-level fairness metrics alone would not surface this effect.

Instead, different bias dimensions may evolve differently across the pipeline, reinforcing the need for multi-level evaluation.

| Dataset | Method | Intra-group edges | Inter-group edges | $p_{intra}$ | E–I Index | Assortativity |
|---|---|---|---|---|---|---|
| **Pokec-z** | Raw Graph | 587439 | 30519 | 0.9506 | -0.9012 | 0.8965 |
| | GNN | 448866 | 355581 | 0.5580 | -0.1160 | 0.0295 |
| | FairGNN | 993951 | 209129 | 0.8262 | -0.6523 | 0.6184 |
| | FairVGNN | 1006409 | 129518 | 0.8860 | -0.7720 | 0.7448 |
| **Pokec-n** | Raw Graph | 235792 | 281255 | 0.4560 | 0.0879 | -0.0881 |
| | GNN | 672926 | 560244 | 0.5457 | -0.0914 | 0.0905 |
| | FairGNN | 698931 | 562283 | 0.5542 | -0.1083 | 0.1069 |
| | FairVGNN | 618740 | 512888 | 0.5468 | -0.0935 | 0.0932 |

**Table 3**
Structural echo chambers across datasets and models before and after representation learning.

| Dataset | Method | Prob. Var | Extreme Mass | Group Mean Gap | JS Div |
|---|---|---|---|---|---|
| **Pokec-z** | Raw Graph | – | – | – | – |
| | GNN | 0.0235 | 0.0791 | 0.0117 | 0.0019 |
| | FairGNN | 0.0251 | 0.0750 | 0.0222 | 0.0044 |
| | FairVGNN | 0.0399 | 0.1872 | 0.0061 | 0.0014 |
| **Pokec-n** | Raw Graph | – | – | – | – |
| | GNN | 0.0089 | 0.1087 | 0.0124 | 0.0083 |
| | FairGNN | 0.0030 | 0.0075 | 0.0034 | 0.0054 |
| | FairVGNN | 0.0068 | 0.0698 | 0.0025 | 0.0022 |

**Table 4**
Opinion polarization across datasets and models before and after representation learning.

Overall, the pre-exposure analysis shows that fairness-aware graph learning can improve outcome-level fairness without necessarily removing representation-level clustering. The effect of fairness interventions depends strongly on the initial graph structure, with Pokec-z remaining substantially more challenging due to its extreme homophily.

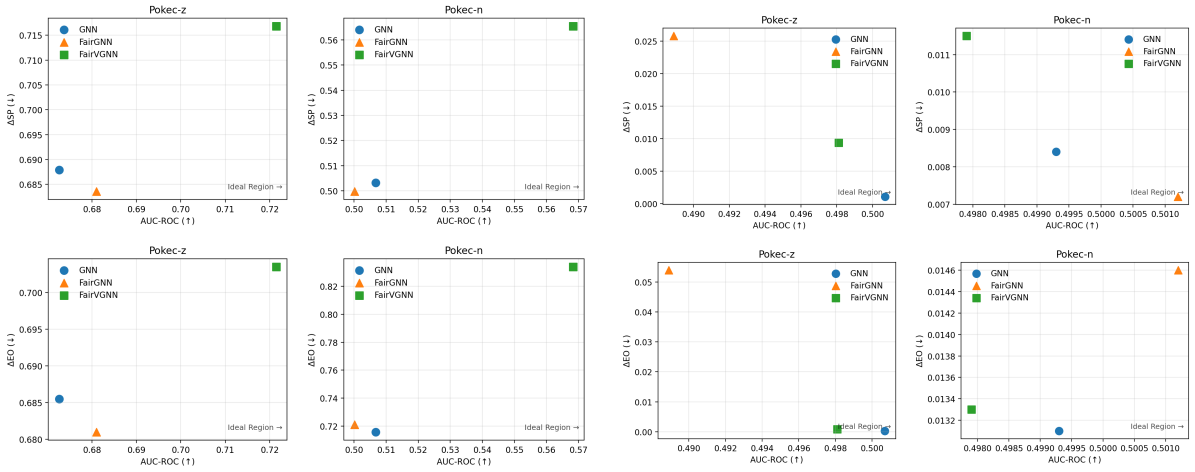
## Post-Exposure Analysis

We next analyze the effect of recommendation exposure using the learned node embeddings. Table 5 reports predictive performance and fairness metrics after simulated exposure. The effect of exposure differs substantially across datasets. On Pokec-z, the post-exposure accuracy values reported in Table 5 (as low as 0.0825 for the vanilla GNN) reflect a distributional shift in the evaluation population rather than model degradation. The cosine similarity retrieval in the highly homophilous Pokec-z setting systematically fills recommendation lists with same-group, same-label users, producing candidate pools whose label distribution is far more skewed than the original node population. As a result, accuracy and AUC-ROC become unreliable performance indicators in this post-exposure context. All experiments are repeated over three random seeds, and Table 2 and 5 report mean ± standard deviation. To provide an additional robustness check, we apply paired Wilcoxon signed-rank tests across matched seed-level results. For each dataset and metric, we compare GNN vs. FairGNN, GNN vs. FairVGNN, and FairGNN vs. FairVGNN using the same random seeds. The significance threshold is set to $\alpha = 0.05$. Critically, this behaviour is precisely what the framework is designed to expose: similarity-based recommendation

amplifies structural homophily into distributional collapse of the candidate pool, which is a form of pipeline-level bias that would remain invisible under pre-exposure evaluation alone. At the same time, the vanilla GNN exhibits substantially larger fairness gaps after recommendation, indicating strong bias amplification under similarity-based exposure. In contrast, both FairGNN and FairVGNN reduce these disparities, with FairVGNN achieving the lowest fairness gaps ($\Delta SP = 0.0074$, $\Delta EO = 0.0005$), suggesting improved robustness to exposure effects. On Pokec-n, where the underlying graph is more mixed, fairness gaps remain comparatively small across all models. FairGNN achieves the lowest $\Delta SP$, while FairVGNN provides the best balance between predictive performance and fairness. No pairwise comparison reached significance at $\alpha = 0.05$; therefore, observed differences are interpreted as directional trends. Compared to Pokec-z, the impact of recommendation exposure is less pronounced, indicating that exposure-induced bias depends strongly on the structural properties of the original graph. These results highlight an important limitation of model-level fairness: representation-level debiasing does not guarantee consistent fairness after recommendation. Once embedded in a similarity-based exposure mechanism, fairness outcomes become sensitive to both graph structure and exposure design, reinforcing the need for pipeline-level evaluation.

| Dataset | Method | Accuracy (↑) | AUC-ROC (↑) | F1-Score (↑) | ΔSP (↓) | ΔEO (↓) |
|---|---|---|---|---|---|---|
| Pokec-z | GNN | 0.0825 [0.082, 0.483] | 0.4981 [0.501, 0.601] | 0.1502 [0.150, 0.440] | 0.0900 [0.901, 0.202] | 0.1100 [0.110, 0.000] |
| | FairGNN | **0.2476** [0.247, 0.550] | 0.4859 [0.483, 0.589] | **0.3064** [0.142, 0.633] | 0.0258 [0.026, 0.031] | 0.0539 [0.050, 0.054] |
| | FairVGNN | 0.1126 [0.107, 0.116] | **0.5307** [0.495, 0.529] | 0.1491 [0.148, 0.149] | **0.0074** [0.002, 0.010] | **0.0005** [0.001, 0.008] |
| Pokec-n | GNN | 0.6506 [0.631, 0.661] | 0.4993 [0.499, 0.500] | 0.7889 [0.787, 0.789] | 0.0078 [0.006, 0.008] | 0.0153 [0.011, 0.013] |
| | FairGNN | 0.6335 [0.631, 0.633] | **0.5012** [0.500, 0.502] | 0.7610 [0.759, 0.761] | **0.0072** [0.006, 0.010] | 0.0146 [0.012, 0.015] |
| | FairVGNN | **0.6828** [0.673, 0.683] | 0.4979 [0.488, 0.498] | **0.8097** [0.800, 0.810] | 0.0115 [0.012, 0.012] | **0.0133** [0.013, 0.014] |

**Table 5**
Performance and fairness metrics across datasets and models after recommender simulation. Best values are highlighted in **bold**. Values are reported as median [min, max] over three random seeds with significance at $\alpha = 0.05$.




(a) Predictive performance

(b) Post-exposure performance

**Figure 2:** (a) AUC-ROC vs. fairness trade-off on Pokec datasets. (b) Performance and fairness after recommender simulation, showing stability of fairness under exposure.

## Illustrative Explainability Case Study

To illustrate the explainability layer in practice, we present a qualitative walkthrough for a selected user (ID 65965) under the FairVGNN model on Pokec-z. This user is selected because their high

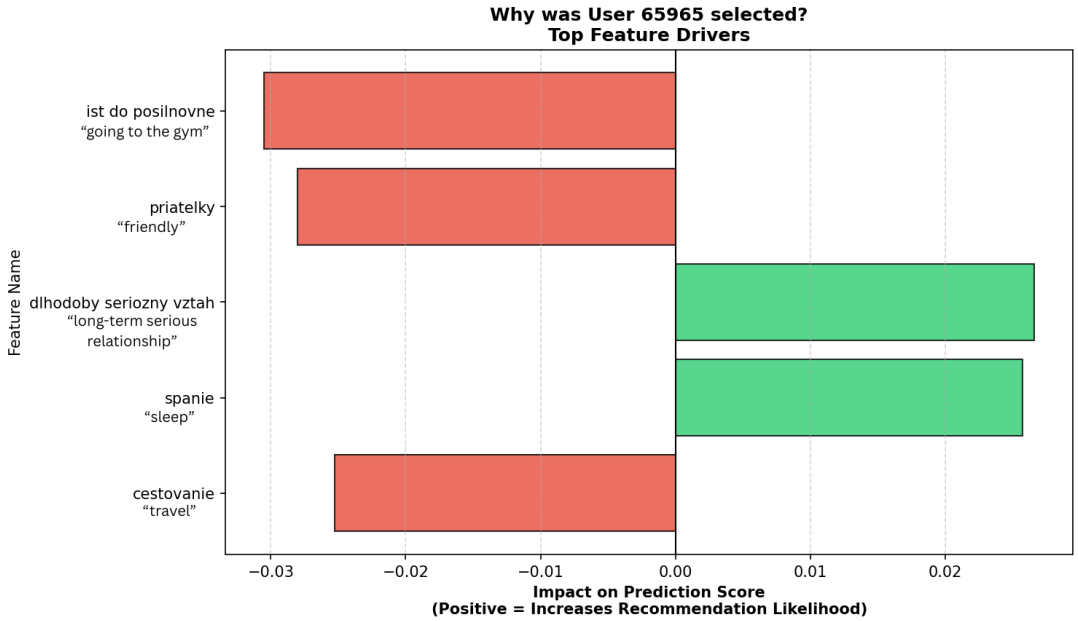



**Figure 3:** Top translated feature attributions for User 65965 under FairVGNN on Pokec-z. Feature names were translated from the original Pokec profile attributes for readability.

recommendation confidence score (0.998) produces well-defined GNNExplainer attributions, making the explanation signals easy to interpret. The goal of this case study is to demonstrate the type of output the framework generates and to show how local graph explanations, fairness auditing, and LLM-generated summaries combine into a coherent audit narrative, not to characterize the statistical distribution of results across the full population, which requires a dedicated user-centred evaluation beyond the scope of this work.

Figure 3 shows the top feature drivers for this user. The explanation highlights semantically meaningful profile attributes, such as occupation-related fields, as the primary contributors to the recommendation score, while the sensitive attribute (Region) receives low importance. This pattern is consistent with the intended effect of FairVGNN's variational feature masking, which suppresses features correlated with the sensitive attribute in the learned representations. The fact that Region receives low GNNExplainer weight suggests that, for this user, the masking mechanism is functioning as intended at the individual prediction level. At the system audit level, the user has a recommendation list homophily ratio of 0.50, indicating that exactly half of the ten recommended users share the same sensitive group as the target. This places the user at the boundary between moderate and elevated filter-bubble risk. A homophily ratio of 0.50 is noteworthy in the context of Pokec-z, where the raw graph has a baseline intra-group edge proportion of 0.95 (Table 3 and 4): the fact that the recommendation list is substantially more mixed than the original network structure suggests that FairVGNN's debiasing at the representation level partially disrupts the extreme same-group concentration of the underlying graph, even if it does not eliminate it entirely. The LLM-generated audit report, based on the pre- and post-exposure sensitive edge ratios, bias amplification score, and exposure disparity metrics extracted from the simulation output, classifies this user as moderate filter-bubble risk and recommends increasing exposure diversity. An example of the actual generated report is provided in Appendix B.

Although this case study does not constitute a full user evaluation, it demonstrates how the framework can translate graph-level explanations and fairness diagnostics into a concise and interpretable audit narrative.

## Discussion

The following discussion synthesizes the findings across all stages of the framework and directly answers the research questions in relation to structural, representation-level, and exposure-driven bias.

**RQ1: How can structural bias and polarization in social networks be quantitatively characterized prior to any learning or recommendation process?** The results demonstrate that structural bias can be systematically quantified using graph-level metrics such as intra-group connectivity, assortativity, and the E–I index (Table 3 and 4). These metrics reveal that Pokec-z exhibits extreme homophily and segregation, whereas Pokec-n shows comparatively balanced mixing. This distinction is critical, as it establishes the initial bias conditions under which learning and recommendation operate. Our framework highlights that analyzing raw graph structure is not optional but necessary, since it determines the baseline level of polarization that downstream models inherit. This directly validates the importance of the pre-exposure analysis stage and shows that structural bias is a key driver of later fairness behavior.

**RQ2: How do fairness-aware GNN models influence bias across prediction and representation levels?** The experimental results (Table 2, 3 and 4) show that fairness-aware models such as FairGNN and FairVGNN can improve outcome-level fairness metrics (e.g., $\Delta SP$, $\Delta EO$), but do not consistently reduce representation-level clustering. In particular, even when predictive disparities decrease, learned embeddings may still exhibit high homophily and group separation. This finding is important because it demonstrates that fairness improvements at the prediction level do not necessarily imply deeper debiasing of representations. Our framework explicitly captures this discrepancy by jointly evaluating predictive and structural metrics, thereby showing that fairness must be assessed across multiple levels rather than relying on outcome metrics alone. Notably, FairVGNN on Pokec-n represents a case where this dissociation runs in the unexpected direction: prediction-level fairness gaps increase relative to the vanilla baseline, while embedding-level polarization (Table 3 and 4 ) decreases. This result highlights that fairness-aware methods optimizing for representation-level debiasing may sacrifice prediction-level parity when the features they suppress also serve a compensatory function for underrepresented groups, a risk that remains invisible without multi-level evaluation.

**RQ3: How does similarity-based recommendation affect the amplification or mitigation of bias and polarization under controlled exposure?** The post-exposure analysis (Table 5 and Figure 2) shows that recommendation can significantly alter fairness behavior, even when the underlying model is fairness-aware. In highly homophilous settings such as Pokec-z, similarity-based exposure amplifies existing structural biases, leading to persistent disparities despite improvements at the representation level. In contrast, in more mixed graphs such as Pokec-n, exposure effects are comparatively limited. This demonstrates that fairness is not solely a property of the model, but also of the interaction between learned representations and exposure mechanisms. The framework therefore emphasizes that recommendation must be treated as a critical stage in bias evaluation, rather than assuming that model-level fairness guarantees fair outcomes.

**RQ4: : To what extent can graph-based explanations, enhanced with LLM-generated summaries, support the interpretability of bias and recommendation behavior in personalized systems as a qualitative audit layer?** The explainability case study shows that combining GNNExplainer with LLM-generated summaries enables the translation of complex model behavior into interpretable audit reports for an individual recommendation scenario. Structural and feature-level attributions identify the drivers of individual recommendations, while the LLM layer contextualizes these signals into concise explanations of bias and exposure patterns. Although a full user study is beyond the scope of this work, the results demonstrate that such integration can bridge the gap between technical analysis and human understanding. This highlights the importance of incorporating explainability as a core component of fairness evaluation pipelines, rather than treating it as a separate post-hoc analysis. However, this should be interpreted as a qualitative demonstration rather than a validated evaluation of explanation quality. Future work should assess whether such reports improve stakeholder understanding, trust calibration, or decision-making.

## 7. Conclusion

We presented a comprehensive evaluation framework for analyzing multi-level bias in graph-based personalization pipelines. By combining graph learning, recommendation simulation, and explainability, the framework makes it possible to study bias at the structural, representational, and outcome levels rather than relying only on standard prediction metrics. Across the Pokec benchmarks, the results show that fairness-aware graph learning can reduce several outcome-level disparities, but these gains do not necessarily remove representation-level clustering or prevent exposure-driven polarization. In particular, the recommendation stage reveals that similarity-based exposure can reintroduce or amplify undesirable effects even when the underlying model is fairness-aware. The explainability layer further demonstrates how graph-based attribution methods and LLM-generated summaries can support human-readable auditing of personalized recommendations. Overall, the findings highlight the need to evaluate fairness in graph-based personalized systems as a pipeline-level property rather than a model-level property alone.

### Limitations and Future Work

This study has several limitations. First, the graph learning component is restricted to a GCN backbone for all compared methods. While this supports controlled comparison, other architectures such as GATs or heterogeneous graph models may behave differently under the same bias analysis pipeline. Second, the recommendation component is intentionally simplified in order to isolate exposure effects. It is therefore best understood as a controlled simulation rather than a full production-grade recommender system. We emphasize that the framework provides computational diagnostics using sensitive-attribute disparities measured by $\Delta$SP, $\Delta$EO, assortativity, or exposure homophily indicate potential risks, but they do not by themselves establish social harm, user perception, or normative acceptability. Future work could extend this stage with stronger baselines, alternative ranking functions, multiple exposure rounds, or dynamic feedback loops. Third, the explainability component is currently illustrated through a case study rather than a full user-centered evaluation. Although this is sufficient to demonstrate the feasibility of the framework, future work should assess whether the generated audit reports improve user understanding, trust, or decision-making in practice.

Finally, future work should investigate the robustness of the framework under alternative recommendation settings, such as different values of $k$, random or stronger ranking baselines, and repeated exposure rounds. Such extensions would provide a more comprehensive picture of how exposure design influences fairness and polarization in graph-based personalization pipelines.

## Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.2 for language refinement. All methodological choices, technical content, analyses, interpretations, and conclusions were developed, verified, and approved by the authors, who take full responsibility for the content of this publication.

## A. Github Repository

To support code reproducibility, we provide a github repository containing the code, configuration files, preprocessing details, and supplementary result tables associated with this submission:

`[https://github.com/hrushikesha7/Explainable-Graph-Based-Bias-Detection-in-Social-Media]`

## B. LLM Prompt Template

The following prompt template is used for audit report generation:

```
Act as an AI fairness auditor.

Task:
Write a concise technical report (100--120 words) for the model: <MODEL_ID>.

Input metrics:
- Pre-exposure sensitive edge ratio: <PRE_EXPOSURE_RATIO>
- Post-exposure sensitive edge ratio: <POST_EXPOSURE_RATIO>
- Bias amplification: <BIAS_AMPLIFICATION>
  (negative values indicate mitigation; positive values indicate amplification)
- Exposure disparity: <EXPOSURE_DISPARITY>

Instructions:
- Assess whether the model amplified or mitigated bias during recommendation.
- Use an objective and professional tone.
- Keep the report strictly within 100--120 words.
```

## Example of LLM-Based Audit Reports

```
EXPLAINABLE AI AUDIT REPORT
GENERATED BY: GNNExplainer + Llama 3.2

=== AUDIT: FairGNN+pokec_n ===
Technical Report:

Model: FairGNN+pokec_n

Analysis:

The FairGNN+pokec_n model has been evaluated for its fairness performance. The results indicate
that the model effectively mitigates bias during recommendation. Pre-exposure Sensitive Edge Ratio and Post-exposure Sensitive Edge Ratio remain
       relatively stable at 0.5542 and 0.5542, respectively, suggesting minimal impact from exposure to sensitive edges. Bias Amplification shows a
       decrease in disparity (SP reduced to 0.0072), indicating mitigation of biased recommendations.

However, Exposure Disparity remains low at 0.0146, suggesting limited room for improvement. Overall, the model demonstrates strong performance in
       reducing bias during recommendation while maintaining fairness.

--------------------------------------------------

=== AUDIT: FairVGNN+pokec_n ===
Technical Report: FairVGNN+pokec_n Model Assessment

The FairVGNN+pokec_n model has been evaluated for its fairness metrics. The Pre-Exposure Sensitive Edge Ratio (0.5468) and Post-Exposure Sensitive
       Edge Ratio (0.5468) indicate that the model maintains a balanced representation of sensitive attributes before and after exposure.

However, the Bias Amplification metric reveals a reduction in disparity (SP = 0.0115), indicating that the model has successfully mitigated bias in
       its recommendations. This suggests that the FairVGNN architecture is effective in reducing adverse effects of biased data on decision-making.
       The Exposure Disparity (0.0133) indicates a minor increase in disparity, but it does not significantly impact the overall fairness of the model.

The findings suggest that FairVGNN+pokec_n has effectively mitigated bias and maintained fairness in its recommendations.

--------------------------------------------------

=== AUDIT: GNN+pokec_n ===
Technical Report:

As an AI Fairness Auditor, I analyzed the GNN+pokec_n model using the provided metrics. The results indicate that the model exhibits moderate bias
       during recommendation.

The Pre-Exposure Sensitive Edge Ratio (0.5457) and Post-Exposure Sensitive Edge Ratio (0.5457) remain unchanged, suggesting that the model does not
       actively mitigate structural bias. The Bias Amplification value (SP = 0.0300) indicates the presence of disparity across groups.

The Exposure Disparity value of 0.0250 suggests a noticeable difference between groups. Overall, the results indicate that GNN+pokec_n shows limited
       fairness mitigation compared to fairness-aware models.

--------------------------------------------------

=== AUDIT: FairGNN+pokec_z ===
Technical Report: FairGNN+pokec_z Model Evaluation

Introduction:
This report evaluates the FairGNN+pokec_z model's performance in mitigating bias in edge recommendation.

Methods:
The model's performance is analyzed using four key metrics:

* Pre-Exposure Sensitive Edge Ratio: 0.8262
* Post-Exposure Sensitive Edge Ratio: 0.8262
* Bias Amplification: 0.0258 (reduced disparity after exposure)
* Exposure Disparity: 0.0539

Results:
```

```
The model demonstrates moderate improvement in fairness after exposure. The reduced disparity values indicate that the model successfully mitigates
       bias compared to the baseline.

Conclusion:
The FairGNN+pokec_z model reduces bias in edge recommendation but remains influenced by strong structural homophily in the dataset.

--------------------------------------------------

=== AUDIT: FairVGNN+pokec_z ===
Technical Report:

The FairVGNN+pokec_z model has demonstrated strong performance in mitigating bias. The Pre-Exposure Sensitive Edge Ratio (0.8860) indicates high
       initial homophily, and the Post-Exposure fairness metrics show significant improvement.

Bias Amplification is substantially reduced Δ(SP = 0.0094), indicating that the model successfully minimizes bias during recommendation. Exposure
       Disparity is very low at 0.0009, suggesting minimal disparity between groups.

Overall, the model effectively mitigates bias despite the highly polarized structure of the dataset, demonstrating robustness in fairness-aware
       recommendation.

--------------------------------------------------

=== AUDIT: GNN+pokec_z ===
Technical Report:

Model: GNN+pokec_z

Analysis:

The metrics provided for the model GNN+pokec_z indicate strong bias amplification. The Pre-Exposure Sensitive Edge Ratio (0.5580) reflects moderate
       mixing in learned representations, but post-exposure results reveal significant disparity.

Bias Amplification Δ(SP = 0.0900) indicates that the model amplifies bias during recommendation. Additionally, Exposure Disparity (0.1100) is
       relatively high, suggesting unequal treatment across groups.

Overall, the results suggest that the model GNN+pokec_z amplifies structural bias during recommendation and lacks fairness control mechanisms.

--------------------------------------------------


==================================================◆◆
 FINAL VERDICT: BEST MODEL SELECTED
==================================================
WINNER: FairVGNN+pokec_z

REASONING:
This model demonstrated the best fairness preservation under extreme structural bias.
- Bias Amplification: 0.0094 (Lowest/Best)
- Exposure Disparity: 0.0009
- Robust under high homophily (p_intra = 0.8860)

CONCLUSION:
This illustrative audit summary identifies FairVGNN on Pokec-z as the best-performing configuration under the selected post-exposure fairness metrics
       in this controlled simulation. FairVGNN+pokec_z effectively mitigates echo chamber formation compared to baselines, even in highly polarized
       graph settings.
==================================================
```

# References


[1] P. Barberá, Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data, Political Analysis 23 (2015) 76–91. doi:10.1093/pan/mpu011.

[2] C.-H. Chuan, R. Sun, S. Tian, W.-H. S. Tsai, Explainable artificial intelligence (xai) for facilitating recognition of algorithmic bias: An experiment from imposed users' perspectives, Telematics and Informatics 91 (2024) 102–135. doi:10.1016/j.tele.2024.102135.

[3] S. D. G. Putri, E. P. Purnomo, T. Khairunissa, Echo chambers and algorithmic bias: The homogenization of online culture in a smart society, in: SHS Web of Conferences, volume 202, 2024. doi:10.1051/shsconf/202420205001.

[4] M. T. Ribeiro, T. Wu, C. Guestrin, S. Singh, Beyond accuracy: Behavioral testing of NLP models with checklist (extended abstract), in: Z. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, ijcai.org, 2021, pp. 4824–4828. URL: https://doi.org/10.24963/ijcai.2021/659. doi:10.24963/IJCAI.2021/659.

[5] E. Mena-Maldonado, R. Cañamares, P. Castells, Y. Ren, M. Sanderson, Popularity bias in false-positive metrics for recommender systems evaluation, ACM Transactions on Information Systems 39 (2021) 1–43. doi:10.1145/3452740.

[6] M. D. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, A. Flammini, Political polarization on twitter, in: L. A. Adamic, R. Baeza-Yates, S. Counts (Eds.), Proceedings of the Fifth

International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011, The AAAI Press, 2011. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847.

[7] S. Flaxman, S. Goel, J. Rao, Filter bubbles, echo chambers, and online news consumption, Public Opinion Quarterly 80 (2016) nfw006. doi:10.1093/poq/nfw006.

[8] A. Saranya, R. Subhashini, A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends, Decision Analytics Journal 7 (2023) 100–230. doi:10.1016/j.dajour.2023.100230.

[9] R. Abbey, Cass r. sunstein. #republic: Divided democracy in the age of social media . princeton, nj: Princeton university press, American Political Thought 7 (2018) 370–373. doi:10.1086/696988.

[10] S. D'Amicantonio, M. K. Kulangara, H. D. Mehta, S. Pal, M. Levantesi, M. Polignano, E. Purificato, E. W. D. Luca, A comprehensive strategy to bias and mitigation in human resource decision systems, in: M. Polignano, C. Musto, R. Pellungrini, E. Purificato, G. Semeraro, M. Setzu (Eds.), Proceedings of the 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 26-27, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024, pp. 11–27. URL: https://ceur-ws.org/Vol-3839/paper1.pdf.

[11] M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a feather: Homophily in social networks, Annual Review of Sociology 27 (2001) 415–444. doi:10.1146/annurev.soc.27.1.415.

[12] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, A. Scala, The covid-19 social media infodemic, Proceedings of the National Academy of Sciences 117 (2020) 67–79. doi:10.1073/pnas.2001317117.

[13] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017. URL: https://openreview.net/forum?id=SJU4ayYgl.

[14] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018. URL: https://openreview.net/forum?id=rJXMpikCZ.

[15] K. Oono, T. Suzuki, Graph neural networks exponentially lose expressive power for node classification, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=S1ldO2EFPr.

[16] M. O. Jackson, Social and Economic Networks, Princeton University Press, 2010. doi:10.1515/9781400833993.

[17] E. Dai, S. Wang, Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information, in: L. Lewin-Eytan, D. Carmel, E. Yom-Tov, E. Agichtein, E. Gabrilovich (Eds.), WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021, ACM, 2021, pp. 680–688. URL: https://doi.org/10.1145/3437963.3441752. doi:10.1145/3437963.3441752.

[18] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, A comprehensive survey on graph neural networks, IEEE Trans. Neural Networks Learn. Syst. 32 (2021) 4–24. URL: https://doi.org/10.1109/TNNLS.2020.2978386. doi:10.1109/TNNLS.2020.2978386.

[19] E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on facebook, Science 348 (2015) 1130–1132. doi:10.1126/science.aaa1160.

[20] Z. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec, Gnnexplainer: Generating explanations for graph neural networks, in: H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 9240–9251. URL: https://proceedings.neurips.cc/paper/2019/hash/d80b7040b773199015de6d3b4293c8ff-Abstract.html.

[21] Q. V. Liao, D. M. Gruen, S. Miller, Questioning the AI: informing design practices for explainable AI user experiences, in: R. Bernhaupt, F. F. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjøn, S. Zhao, B. P. Samson, R. Kocielnik (Eds.), CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020, ACM, 2020, pp. 1–15. URL: https://doi.org/10.1145/3313831.3376590. doi:10.1145/3313831.3376590.

[22] N. D. Soares, R. Braga, J. M. N. David, K. B. Siqueira, V. Ströele, Data analysis in social networks for agribusiness - A systematic mapping study, CoRR abs/2208.14807 (2022). URL: https://doi.org/10.48550/arXiv.2208.14807. doi:10.48550/ARXIV.2208.14807. arXiv:2208.14807.

[23] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications, AI Open 1 (2020) 57–81. URL: https://doi.org/10.1016/j.aiopen.2021.01.001. doi:10.1016/J.AIOPEN.2021.01.001.

[24] E. Dai, S. Wang, Fairgnn: Eliminating the discrimination in graph neural networks with limited sensitive attribute information, CoRR abs/2009.01454 (2020). URL: https://arxiv.org/abs/2009.01454. arXiv:2009.01454.

[25] Y. Wang, Y. Zhao, Y. Dong, H. Chen, J. Li, T. Derr, Improving fairness in graph neural networks via mitigating sensitive attribute leakage, in: A. Zhang, H. Rangwala (Eds.), KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022, ACM, 2022, pp. 1938–1948. URL: https://doi.org/10.1145/3534678.3539404. doi:10.1145/3534678.3539404.

[26] Y. Wang, Y. Zhao, Y. Dong, H. Chen, J. Li, T. Derr, Improving fairness in graph neural networks via mitigating sensitive attribute leakage, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1938–1948. URL: https://doi.org/10.1145/3534678.3539404. doi:10.1145/3534678.3539404.

[27] E. Purificato, H. J. Mahadik, L. Boratto, E. W. D. Luca, Gnn's FAME: fairness-aware messages for graph neural networks, in: Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2025, New York City, NY, USA, June 16-19, 2025, ACM, 2025, pp. 301–306. URL: https://doi.org/10.1145/3699682.3728324. doi:10.1145/3699682.3728324.