Fairness-Accuracy Trade-offs in Group Recommendations: A Nash Bargaining Approach

Tazkia Binty Faruqui[0009−0004−1238−0888], Maria Stratigi[0000−0003−2482−4605], and Kostas Stefanidis[0000−0003−1317−8062]

Data Science Research Centre, Tampere University, Finland
faruqui.586@gmail.com, {maria.stratigi,konstantinos.stefanidis}@tuni.fi

Abstract. Group recommender systems (GRSs) must reconcile the competing objectives of fairness, ensuring equitable satisfaction among group members, and accuracy, delivering recommendations that closely align with collective preferences. Traditional approaches often assume a trade-off between fairness and accuracy. This paper investigates whether a Nash Bargaining-based framework can achieve high levels of both objectives under the evaluated experimental setting. The Nash Bargaining Solution (NBS) is employed to generate Pareto-efficient, symmetric, and individually rational group outcomes that guarantee balanced satisfaction across all members. Using the MovieLens 10M dataset, the study constructs genre- and tag-based user subgroups to evaluate fairness behavior across both macro- and micro-level content categorizations. The resulting fair group recommendations form the baseline for a large-scale counterfactual validation framework that assesses robustness under systematic catalog perturbations. Empirical results suggest that the proposed framework can maintain high recommendation accuracy while preserving stable fairness under the evaluated counterfactual scenarios.

Keywords: Group Recommender Systems · Nash Bargaining Theory · Fairness–Accuracy Trade-off · Counterfactual Analysis.

1 Introduction

Designing effective *Group Recommender Systems* (GRSs) is a challenging task because it must balance two competing objectives: *fairness* and *accuracy*. Fairness ensures that the satisfaction of all group members is equitably considered, preventing domination by any single user, whereas accuracy measures how closely the recommended item aligns with the group's preferences. Traditional aggregation approaches prioritize accuracy but often neglect fairness, leading to biased outcomes. Conversely, methods that attempt to enforce fairness constraints tend to reduce predictive accuracy, producing less relevant recommendations. This tension has traditionally been interpreted as a fairness–accuracy trade-off in group recommendation. However, recent theoretical advances suggest that when fairness is treated as a cooperative optimization objective rather than a constraint, both fairness and accuracy can be jointly improved. This paper provides empirical evidence suggesting that fairness and accuracy can be jointly maintained under the evaluated setting.

The *Nash Bargaining Solution* (NBS), a key concept in cooperative game theory, provides a mathematically grounded method for achieving fair outcomes among multiple agents. It does so by maximizing the product of users' utility gains relative to their disagreement points, ensuring *Pareto efficiency*, *symmetry*, and *individual rationality*. When applied to GRSs, NBS guarantees that every member benefits proportionally, enabling fairness to emerge organically from the optimization process rather than being imposed externally. Despite this potential, the use of Nash Bargaining in group recommendations remains underexplored, particularly under dynamic catalog conditions where content availability or user preferences can change. This paper addresses this research gap by combining Nash Bargaining optimization with a counterfactual robustness analysis framework to evaluate fairness stability under controlled catalog perturbations.

Specifically, this paper proposes a solution grounded in Nash Bargaining Theory, which treats fairness as a cooperative optimization objective rather than a constraint. In this setting, each group member is modeled as a rational agent, and the collective recommendation is determined by maximizing the Nash product of their utility deviations from the group's disagreement point (mean utility). This guarantees equitable allocation of benefits while preserving group satisfaction.

However, fairness must not only be achieved but also proven stable under change. To that end, this research introduces a dual-granularity counterfactual analysis framework that systematically removes genres (macro-level) and tags (micro-level) from the movie catalog and recomputes recommendations to measure shifts in fairness and accuracy. This approach enables the identification of influential content subgroups that most strongly affect group utility balance. By combining game-theoretic fairness with counterfactual validation, we provide an interpretable understanding of fairness behavior in group recommendations.

In summary, this work makes the following contributions:

- Introduces a cooperative game-theoretic model that ensures equitable satisfaction and Pareto efficiency, enabling fairness to emerge naturally without post-hoc correction.
- Proposes a novel counterfactual sensitivity analysis framework that examines fairness stability at both macro (genre) and micro (tag) levels through extensive catalog perturbations, analyzing all genres and the most frequently occurring tags.
- Demonstrates the framework's performance using a large-scale dataset, forming 1,000 user groups and conducting 119 counterfactual experiments involving 19 genres and 100 user-generated tags.
- Provides empirical evidence that fairness and accuracy exhibit a positive association within the evaluated setting in group recommendations when fairness is optimized cooperatively via Nash Bargaining.

2 Related Work

Fairness in Recommendation Systems. Fairness-aware recommendation systems have evolved from single-objective optimization toward multi-stakeholder and multi-objective formulations. Early approaches introduce fairness constraints within model learning. For example, [3] proposes F2VAE, incorporating fairness in both representation learning and recommendation generation, while [8] addresses fairness and diversity in sequential group recommendations, highlighting trade-offs across recommendation rounds. Similarly, [5] integrates fairness with interactivity and explanations, emphasizing transparency and user trust. A broader overview of fairness definitions and methodological directions is provided in [12]. User-centric fairness has also received significant attention. [9] demonstrates that collaborative filtering methods tend to favor active users, proposing re-ranking strategies to balance utility across user groups. Extending this idea, [13] introduces FaiREO, a multi-objective optimization framework that balances fairness and accuracy with minimal performance loss. [1] links fairness issues to popularity bias, showing how overexposure of popular items leads to systematic unfairness. More recently, [7] proposes an energy-based framework capturing distributional disparities across users and items. Overall, these works highlight a shift toward dynamic, distribution-aware, and explainable fairness mechanisms that attempt to mitigate bias while preserving recommendation quality.

Nash Bargaining in Recommender Systems. Game-theoretic formulations provide principled mechanisms for balancing individual and collective utility in group recommendation. [22] introduces one of the earliest Nash equilibrium-based approaches, modeling group preference aggregation as a strategic interaction problem. Their matrix factorization framework enabled consensus recommendations reflecting both individual and group-level preferences. [21] extends this idea to item-group recommendations, incorporating both individual preferences and social influence to achieve stable group decisions. Negotiation-based and multi-agent approaches further explored fairness in group settings. [4] proposes a negotiation framework where agents iteratively reach agreement under fairness and privacy constraints. Similarly, [11] models recommendation as a non-cooperative game, demonstrating improvements in both fairness and accuracy in educational settings. Cooperative game-theoretic methods have also been explored; for instance, [2] uses the Shapley value for user clustering, improving fairness and scalability in collaborative filtering. More directly related to fairness, [18] argues that Nash bargaining provides a theoretically sound alternative to average-utility maximization, as it enforces proportional fairness by maximizing the product of utilities. [15] further extends Nash bargaining with behavioral considerations such as regret, improving robustness in realistic group decision scenarios. Despite these advances, the application of Nash bargaining in dynamic recommendation environments and its empirical validation under changing conditions remain limited.

Counterfactual Methods in Group Recommender Systems. Counterfactual reasoning has emerged as a powerful tool for explainability and fairness evaluation in recommenders. [16] introduces a model-agnostic framework for generating counterfactual explanations in group recommendations, focusing on transparency while preserving privacy. [17] pioneers counterfactual explainable recommendation, identifying minimal changes in item features that alter recommendation outcomes, thus revealing causal relationships. Subsequent works extended counterfactual analysis to more complex settings. [14] incorporates textual and tabular features to generate interpretable explanations, while [19] proposes a fairness-oriented counterfactual framework that evaluates bias through simulated interventions without requiring demographic data. [6] introduces temporal counterfactual explanations for sequential recommendation, ensuring realistic and coherent user behavior modeling and [20] combines rule-based learning with counterfactual reasoning to improve interpretability and efficiency. Overall, counterfactual approaches provide a causal perspective on recommendation behavior, enabling the analysis of robustness, fairness sensitivity, and feature influence. However, their integration with game-theoretic fairness mechanisms, particularly in group recommendation settings, remains underexplored.

3 Methodology

The Nash Bargaining approach provides a foundation for generating fair and Pareto-optimal group recommendations. This section formally defines the Nash bargaining process, presents its mathematical derivation, describes the algorithmic implementation, and outlines validation and complexity analysis. Formally, given a group of n users $G = \{u_1, u_2, \ldots, u_n\}$ and a set of movies $M = \{m_1, m_2, \ldots, m_k\}$, the Nash Bargaining Solution (NBS) seeks the movie $m^* \in M$ that maximizes the product of each user's utility gain over their disagreement (fallback) utility. Each user u_i has a utility function $u_i(m)$ representing their normalized satisfaction for movie m, where $0 \leq u_i(m) \leq 1$.

Individual Disagreement Point: $d_i = \frac{1}{|M_i|} \sum_{m \in M_i} r_{i,m}$, where M_i is the set of movies rated by user i, and $r_{i,m}$ is the rating assigned by u_i to movie m. The disagreement point d_i represents the expected satisfaction if no agreement is reached.

Nash Product to Maximize: $m^* = \underset{m \in M}{\arg\max} \ \prod_{i=1}^{n} \big(u_i(m) - d_i\big)$. The corresponding Nash utility product is: $U_{\text{Nash}} = \prod_{i=1}^{n} \big(u_i(m^*) - d_i\big)$ [10]. This formulation ensures: (i) Pareto efficiency: No other movie improves one user's utility without harming another, (ii) Symmetry: Users with identical preferences receive identical treatment, (iii) Invariance to affine transformations: Scaling of utilities does not affect the outcome, and (iv) Individual rationality: No user receives less than their disagreement point.

3.1 Subgroup Formation

User subgroups were created to enable fairness-oriented analysis across diverse user communities. Two subgrouping strategies were implemented: genre-based subgrouping and tag-based subgrouping.

Genre-Based Subgroups The genre field in the movie metadata frequently contains multiple genre labels for a single movie. These labels were separated and expanded into individual entries, enabling each movie to be associated with multiple genres. The ratings data were then linked with the expanded genre mappings to compute the number of movies rated by each user within each genre. To identify highly active users, the distribution of user-genre rating counts was analyzed, and the top 20% of users within each genre were selected. These users formed the genre-based subgroups, representing communities with strong interest and activity in specific content categories.

Tag-Based Subgroups The tag data were first cleaned by removing missing or blank tags. The frequency of tag usage for each user was then computed to quantify engagement with specific tagging patterns. Users whose tagging activity exceeded the 80th percentile were identified as highly engaged and grouped into tag-based subgroups.

This quantile-based filtering ensured that the resulting subgroups reflected meaningful engagement patterns while reducing noise from infrequent user activity.

3.2 Algorithmic Implementation.

The algorithm computes group recommendations in four major steps: (1) **Group Formation:** Random fixed-size user groups were formed within genre-based subgroups to balance diversity and computational feasibility. Since MovieLens does not provide naturally occurring user groups, the groups used in this study are synthetically constructed through sampling. A sampling strategy was used to limit combinations while preserving representativeness. (2) **Utility Computation:** User utilities were estimated using the trained Surprise SVD collaborative filtering model. Since the MovieLens dataset is sparse, the model predicts user–movie ratings for candidate movies, allowing utilities to be obtained even when explicit ratings are unavailable. The resulting utility estimates, therefore, depend on the predictive quality of the trained SVD model. These predicted utilities were then used in the Nash bargaining computation. (3) **Nash Product Optimization:** For each candidate movie, user utilities were obtained from the collaborative filtering model, and the Nash product was computed as: $P(m) = \prod_{i=1}^{n} \big(u_i(m) - d_i\big)$. The optimal recommendation was then selected as: $m^* = \arg\max_{m \in M} P(m)$. (4) **Output Generation:** For each group, the selected movie m^* and its utility product U_{Nash} were stored as the Nash-optimal recommendation. These values served as the baseline for subsequent counterfactual and fairness–accuracy analyses.

Validation of Nash Bargaining Axioms: The following validations were performed to confirm theoretical correctness: **Pareto check:** U_{Nash} did not increase under any movie switch that reduced any $u_i(m)$; **Symmetry test:** Swapping identical utility vectors yielded identical m^*; **Affine invariance test:** Linear scaling of ratings did not alter the recommendation outcome; **Individual ratio-**

nality: Verified that $u_i(m^*) \geq d_i$ for all users i; **Fairness sensitivity:** Observed a consistent balance between utility balance and collective satisfaction.

Computational Complexity. Let n be the average group size, M_G denote the number of candidate movies evaluated for each group, and G the total number of groups. The time complexity is: $O(G \cdot M_G \cdot n)$, scaling linearly with the number of groups, candidate movies, and users per group.

4 Counterfactual Analysis Framework

This phase builds upon the baseline recommendations produced by Nash bargaining. It perturbs the movie catalog to quantify how the removal of certain genres or tags influences group-level utilities and fairness outcomes. By comparing the perturbed (counterfactual) outcomes with the baseline, the analysis identifies which content attributes have the greatest influence on group decisions. Each simulation follows the same principle: (i) Identify a subgroup of movies defined by a shared attribute (e.g., genre or tag), (ii) Remove all movies belonging to that subgroup from the catalog, (iii) Recompute the Nash bargaining solution for all user groups using the reduced catalog, and (iv) Compare baseline and counterfactual recommendations to measure the magnitude of influence. In this study, robustness refers to the stability of recommendation outcomes under catalog perturbations, while counterfactual sensitivity measures the changes in utility or recommendations after genre or tag removal. These controlled removals assess sensitivity rather than establish causal effects.

Genre-Based Counterfactual Analysis. In the genre-based setting, each genre is iteratively removed from the catalog, and the Nash bargaining process is recomputed for all user groups. For a given genre γ:

1. Filter the catalog to exclude all movies belonging to γ to obtain $M' = \{\, m \in M : \gamma \notin \text{genres}(m) \,\}$.
2. Recompute the group recommendation process and utility product on M'.
3. Compute the influence scores and the recommendation change rate.

This provides a macro-level view of how broad content categories affect fairness and satisfaction in group decisions.

Tag-Based Counterfactual Analysis. Complementing the genre-level study, tag-based analysis focuses on user-generated tags that capture nuanced characteristics of movies. For a frequent tag τ:

1. Filter the catalog to exclude all movies with tag τ to obtain $M' = \{\, m \in M : \tau \notin \text{tags}(m) \,\}$.
2. Recompute the group recommendation process and utility product on M'.
3. Compute the influence scores and the recommendation change rate.

This enables finer-grained sensitivity profiling, revealing how specific themes or stylistic attributes drive changes in group recommendations.

5 Comparative Analysis

This framework integrates results from the baseline Nash bargaining process and the counterfactual influence analyses to measure the extent of recommendation changes and utility differences after specific subgroups were removed from the catalog.

Genre-based Comparative Analysis. This phase merged baseline group recommendations with genre-level counterfactual outcomes.

Change Rate. For each removed genre subgroup, the framework calculated the proportion of groups whose top recommendation changed: Change Rate $= \frac{N_{\text{changed}}}{N_{\text{total}}}$, where N_{changed} represents the number of groups whose recommended movie differed from the baseline, and N_{total} is the total number of groups analyzed. This metric quantified the stability of recommendations when specific genres were excluded.

Influence Summary. The influence of each removed genre subgroup was summarized statistically by computing the mean, median, and maximum of the influence score distribution: Influence Score $= |U_{\text{baseline}} - U_{\text{counterfactual}}|$. U_{baseline} is the utility product computed on the original catalog, and $U_{\text{counterfactual}}$ is the utility product computed after removing the subgroup. These values were aggregated across all counterfactual scenarios to identify the most influential genres contributing to fairness and utility changes.

Utility Comparison. For every subgroup, the average baseline utility and counterfactual utility were compared to estimate relative performance differences: Relative Drop (%) $= \frac{U_{\text{baseline}} - U_{\text{counterfactual}}}{U_{\text{baseline}}} \times 100$. This measure indicated how strongly each genre removal affected group satisfaction levels.

Tag-Based Comparative Analysis. The tag-based analysis provided micro-level insights into stability and sensitivity. It compared baseline utilities with tag-based counterfactuals to identify fine-grained sources of variation.

Utility Comparison. For each tag, baseline and counterfactual utilities were compared to visualize deviation from equilibrium. A diagonal line ($y = x$) was plotted as a reference of no change, and dispersion from this line indicated the extent of influence caused by tag removals.

Influence Distribution. The influence score was calculated as: Influence Score $= |U_{\text{baseline}} - U_{\text{counterfactual}}|$. These visualizations revealed the tags that caused the largest changes in group fairness and satisfaction.

Change Rate by Tag. The recommendation stability for each tag was computed using Change Rate $= \frac{N_{\text{changed}}}{N_{\text{total}}}$, where N_{changed} is the count of groups whose recommendations changed after removing movies with a given tag. The resulting data identified tags that contributed most to recommendation instability.

Top Influential Tags. A ranked bar chart of the top influential tags summarized which content-level features exhibited the strongest counterfactual impact on the system. The tag-level framework also produced a comprehensive dashboard combining influence, change rate, and utility metrics for cross-comparison.

Accuracy Assessment.

Mean True Rating (MTR). The MTR is the average actual rating assigned by group members to the recommended movie: $\text{MTR}(g, m^*) = \frac{1}{|G_{\text{rated}}|} \sum_{u \in G_{\text{rated}}} r_{u,m^*}$, where G_{rated} denotes the subset of group members who rated the recommended movie m^*, and r_{u,m^*} is the rating given by u to m^*. A higher MTR indicates that the recommendation closely matches group preferences.

Precision@1. This relevance metric captures whether the recommended movie meets a defined threshold: Precision@1$(g, m^*) = \mathbb{1}[\,\text{MTR}(g, m^*) \geq \theta\,]$, where $\theta = 4.0$ represents the relevance threshold (ratings of 4.0 or higher are considered "liked"), and $\mathbb{1}[\cdot]$ is an indicator function that equals 1 if the condition holds and 0 otherwise. When aggregated across all groups, Precision@1 represents the proportion of group recommendations that meet the relevance threshold.

6 Fairness–Accuracy Trade-off Analysis

This phase utilized the output of the previous accuracy evaluation to jointly analyze fairness and accuracy metrics. Both dimensions were normalized to a common scale to enable direct comparison and visualization. The framework produced multiple visual representations illustrating their relationship across groups, subgroup levels, and group sizes.

Metric Normalization. To ensure comparability between fairness and accuracy, each metric was scaled into the $[0, 1]$ range using min–max normalization: $X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$, where X represents either the Nash utility (fairness) or the mean true rating (accuracy). The resulting normalized variables, normalized fairness and normalized accuracy, were used for all trade-off analyses.

Fairness–Accuracy Relationship. The core analysis compared normalized fairness and normalized accuracy across all group recommendations using scatter plots and correlation analysis. The Pearson correlation coefficient was computed as $r = \frac{\text{Cov}(F,A)}{\sigma_F\,\sigma_A}$, where F and A represent the normalized fairness and accuracy values, respectively. A high positive correlation indicated that improving fairness did not reduce accuracy.

Binned Fairness Analysis. To further interpret the fairness–accuracy relationship, the normalized fairness scores were divided into quantile-based bins. For each bin B, the framework computed Mean Accuracy $= \frac{1}{|B|} \sum_{i \in B} A_i$ and Mean Precision@1 $= \frac{1}{|B|} \sum_{i \in B} P_i$, where A_i is the normalized accuracy and P_i is the Precision@1 value for instance i. This enabled visualization of how accuracy metrics behave across different fairness levels.

Group Size Analysis. To assess the effect of group size on fairness–accuracy dynamics, each group's size n was identified and color-coded in scatter plots. For groups of the same size n, mean fairness and accuracy were computed as $\bar{F}_n = \frac{1}{|G_n|} \sum_{i \in G_n} F_i$ and $\bar{A}_n = \frac{1}{|G_n|} \sum_{i \in G_n} A_i$, where G_n denotes the set of all groups of size n.

7 Results and Analysis

Nash Bargaining-based Group Recommendations. The Nash Bargaining Solution (NBS) was applied to user groups formed from genre-based subgroups to model collective decision-making. Each fixed-size user group jointly selected a movie that maximized the product of their utility gains relative to the disagreement point, ensuring fairness across all members. The optimized implementation (sampled combinations per genre and memory-efficient computation) enabled large-scale analysis on the MovieLens 10M dataset. This baseline serves as the foundation for all subsequent counterfactual and fairness–accuracy evaluations presented in this section. We first evaluate the stability of Nash-based recommendations under genre-level removals.

Overall Stability Analysis. As shown in Figure 1, there is a strong positive correlation between baseline and counterfactual utilities, confirming that the Nash-based recommendations are robust to catalog changes. The clustering around the diagonal line implies minimal fairness distortion across most groups when genres are selectively removed.

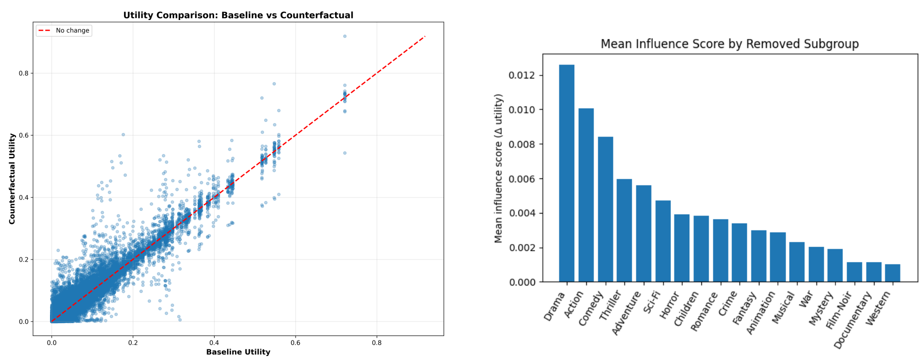


Fig. 1: Genre-level counterfactual results: utility comparison (left) and mean influence score by removed genre (right).

Genre-Specific Influence. Drama, Action, and Comedy shows the highest influence scores, indicating that their removal most strongly alters group utilities and fairness balance. Lower scores for genres like Documentary and Western suggest minimal global impact on group satisfaction. Figure 1 also shows the mean influence score (Δ utility) per removed genre. The results highlight that *Drama*, *Action*, and *Comedy* contribute most to collective fairness, whereas other genres play smaller roles in shaping group consensus.

Recommendation Sensitivity. Drama causes the largest change rate (54.5%), followed by Comedy (38.7%) and Romance (29.2%), reflecting these genres' strong influence on recommendation decisions. As seen in Figure 2, the removal of *Drama* resulted in over half of the recommendations changing, confirming its dominant role in shaping group satisfaction. Genres with lower change rates, such as Documentary or Western, indicate higher resilience of the system's fairness structure.

Utility Impact Analysis. Positive values indicate a utility increase after removal; negative values reflect utility loss. The removal of Comedy slightly increases overall utility, whereas removing Drama and Action decreases group

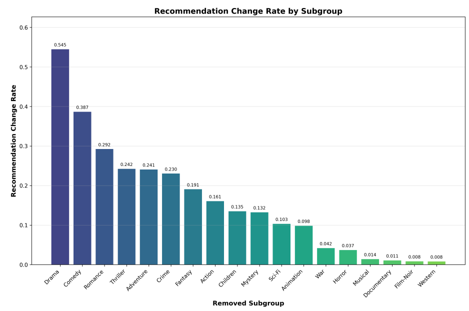


Fig. 2: Recommendation change rate by removing genre.

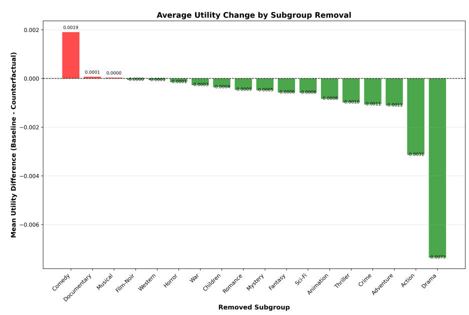


Fig. 3: Average utility difference (baseline − counterfactual) by genre removal.

satisfaction, demonstrating asymmetric genre dependencies. Figure 3 shows how removing certain genres affects the mean group utility. While removing *Comedy* or *Documentary* leads to small positive utility gains, removing *Drama* or *Action* results in noticeable utility reductions, indicating these genres' higher contribution to fairness and user satisfaction.

Counterfactual Validation on Tag Removals. This section extends the counterfactual validation to a finer level of granularity by analyzing user tagging behavior rather than broad genre categories. Tags often capture subtle thematic or stylistic preferences, such as *dark humor*, *strong female lead*, or *based on a true story*, that cannot be represented by genres alone. By removing movies associated with individual tags from the catalog and recomputing Nash Bargaining recommendations, we assess how such micro-level perturbations affect fairness stability, utility, and recommendation behavior.

The goal of this analysis is to evaluate the resilience of the Nash Bargaining framework under fine-grained preference disruptions and to understand how tag-level semantics contribute to fairness sensitivity.

Overall Tag-Level Stability. As shown in Figure 4, each point represents a user group's utility before and after removing movies associated with a particular tag. The majority of points lie densely along the red dashed diagonal, indicating a strong linear relationship between baseline and counterfactual utilities. This close alignment shows that most tag removals cause only minimal deviations

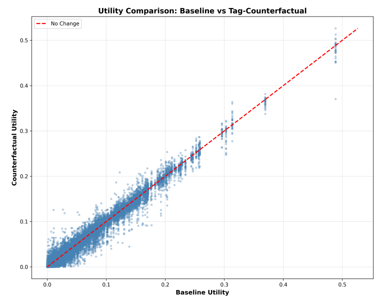


Fig. 4: Utility comparison between baseline and counterfactual Nash utilities at the tag level.

in group utility. A small number of points drift slightly above or below the diagonal—primarily for groups with higher baseline utilities—suggesting that the effect of tag removal is localized rather than widespread. Overall, the tight clustering confirms that the Nash Bargaining framework remains highly stable even when fine-grained, tag-level thematic attributes are perturbed.

Tag Influence Distribution. Figure 5 (right) shows the distribution of

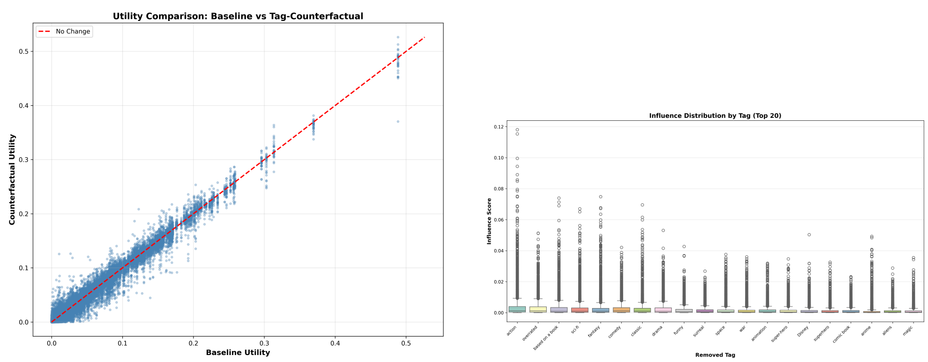


Fig. 5: Tag-level counterfactual results: utility comparison between baseline and counterfactual Nash utilities (left), and distribution of influence scores for the most common tags (right).

influence scores for the top twenty tags. Most tags display very low median influence and compact interquartile range ranges, indicating that their removal has only a minor effect on group utilities. However, a few tags—such as *action*, *overrated*, *based on a book*, *fantasy*, and *drama*—exhibit wider spreads and higher upper whiskers, suggesting that for some user groups these attributes can substantially affect the Nash utility. Toward the right side of the plot, tags like *anime*, *aliens*, and *magic* also generate occasional high outliers even though their central tendency remains small. Overall, the boxplots indicate that fairness effects are generally stable across tags, but the removal of certain thematic attributes can create localized fairness sensitivity for specific subpopulations.

Most Influential Tags. Tags such as *imdb top 250*, *overrated*, *"owned"*, *"dvd"* , *"own"* , and *"classic"* have the strongest impact on group utility and fairness balance, while others like *biographical* and *historical* show minimal effect. The result reveals that emotionally charged or stylistically distinctive tags tend to exert a stronger influence on group satisfaction. These effects remain localized—specific to the groups most aligned with those tags—without destabilizing the overall fairness equilibrium.

Recommendation Change Rate. As shown in Figure 6, higher bars indicate tags whose removal caused a larger proportion of group recommendations to change. This suggests that certain tags act as strong preference signals, significantly influencing group decisions. The tag based on a book produces the highest change rate (21.83%), followed by classic (17.26%) and drama (12.55%), showing that removing these thematic attributes significantly alters the final group decision. Mid-range tags such as sci-fi, action, fantasy, and true story generate moderate shifts, while tags like Pixar, remake, space, and drugs produce only minor effects.

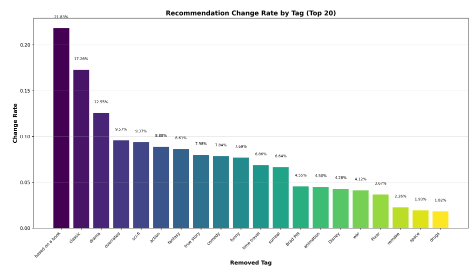


Fig. 6: Recommendation change rate by removing the tag.

The overall pattern confirms that certain tags act as strong preference signals within groups, and their removal can meaningfully influence the recommended item. Tag-level perturbations produced stronger and more consistent fairness impacts than genre-level counterfactual experiments, indicating that fine-grained user-generated tags play a central role in shaping the fairness structure of the Nash Bargaining framework.

Summary of Tag-Based Counterfactual Analysis. The tag-level counterfactual results show that the Nash Bargaining framework remains highly stable under fine-grained semantic perturbations. The utility comparison reveals that most groups retain almost identical utilities after tag removal, with points closely aligned along the no-change diagonal. Influence distributions further indicate that most tags have low and consistent impact, with only a few showing wider variability for specific user communities.

Ranking tags by mean influence highlights attributes such as *action*, *overrated*, and *based on a book* as the most impactful. The recommendation change-rate analysis similarly shows that removing certain tags—especially *based on a book*, *classic*, and *drama*—causes the most noticeable shifts, while others create minimal effects. These shifts further confirm that tag-level removals can produce stronger and more consistent fairness impacts than genre-level removals. Overall, tag-level removals introduce localized changes but do not disrupt the global fairness–accuracy balance, confirming the robustness of the Nash Bargaining model.

Cross-Comparison Summary. A comparative synthesis of the genre- and tag-based counterfactual analyses confirms the global robustness of the Nash Bargaining framework. At the genre level, the removal of dominant categories such as Drama, Action, and Comedy produced noticeable fluctuations in group utilities and recommendation outcomes. However, tag-level removals generated stronger and more consistent fairness impacts, indicating that fine-grained user-generated semantic attributes exert greater influence on Nash bargaining outcomes than broad genre categories.

Together, these results show that fairness stability in Nash-based group recommendations is multi-scale, with fine-grained semantic tag attributes exerting stronger counterfactual influence on fairness outcomes than broader genre-level content categories.

As shown in Figure 7, the combined Influence heatmap highlights that only a limited number of genres and tags exert notable fairness sensitivity. Overall, the findings validate that the proposed Nash Bargaining approach maintains a consistent fairness–accuracy equilibrium across both coarse (genre) and fine-grained (tag) semantic dimensions of the movie catalog.

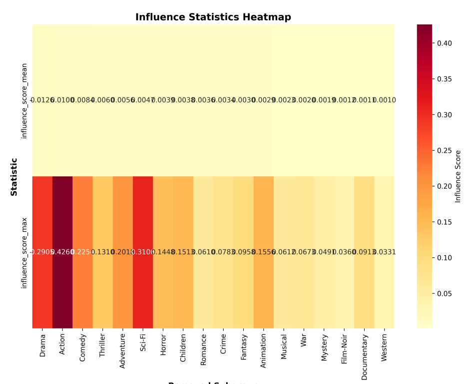


Fig. 7: Combined influence analysis integrating genre- and tag-level counterfactuals. Warmer areas correspond to higher mean and maximum influence values, illustrating that only a few genres and tags exert notable fairness sensitivity.

Accuracy Evaluation. This section evaluates the predictive realism of the recommended items. We first inspect the global distribution of mean true ratings for the recommended movies, followed by an analysis of the success rate (Precision@1) across group sizes.

As shown in Figure 8, the distribution is concentrated at or above the relevance threshold, indicating strong overall accuracy. Accuracy is evaluated across the fixed-size user groups. Figure 9 shows that accuracy remains consistently

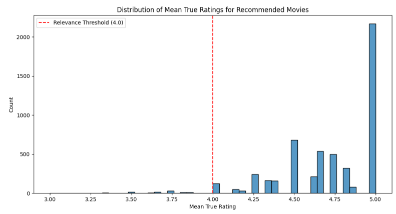


Fig. 8: Distribution of mean true ratings for recommended movies.

high across the evaluated fixed-size user groups. The vertical reference at 4.0 denotes the relevance threshold, and the concentration above this line confirms strong overall accuracy (Mean True Rating = 4.7226, Precision@1 = 98.34%). Overall, accuracy remains consistently high across the evaluated fixed-size user groups; the success rate is consistently high for all groups (Mean True Rating = 4.7226, Precision@1 = 98.34%).

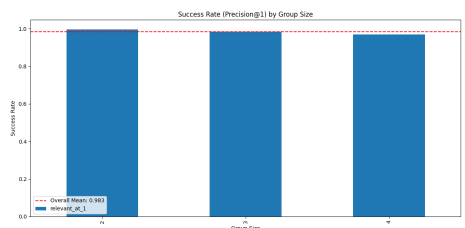


Fig. 9: Success rate (Precision@1 = 98.34%) across evaluated fixed-size user groups with an overall mean reference line. The consistently high bars across sizes show stable performance.

Fairness–Accuracy Trade-off. We analyze how Nash fairness relates to realized accuracy. As shown in Figure 10, the evaluated groups show a positive association between fairness and accuracy. Three complementary visuals are used: a direct correlation view, a progressive view where groups are ordered by fairness, as illustrated in Figure 11a, and a cumulative success curve that reflects stability as fairness increases, as shown in Figure 11b.

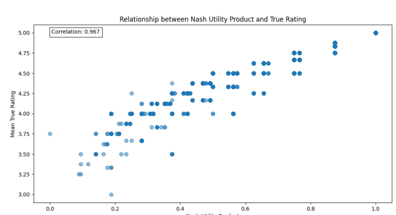
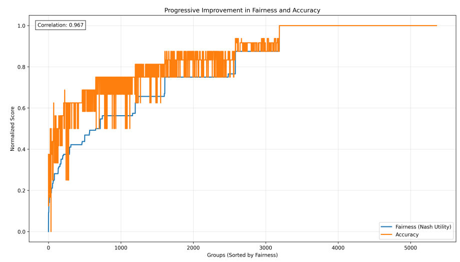
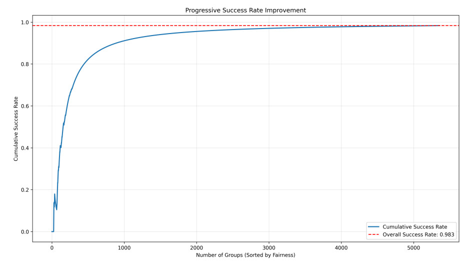


Fig. 10: Relationship between Nash utility product (fairness) and MTR (accuracy). The upward trend indicates a positive association between fairness and accuracy.



(a) Progressive improvement



(b) Cumulative success rate

Fig. 11: Relationship between fairness and recommendation accuracy across group evaluations. The results show that higher fairness values are generally associated with stable or increasing recommendation quality in the evaluated setting.

8 Conclusions

This study investigated whether fairness and accuracy can be jointly optimized in group recommendations by combining NBS with a counterfactual validation framework. Using MovieLens 10M, we evaluated 1,000 fixed-size user groups and conducted 119 counterfactual experiments across 19 genres and 100 user-generated tags. The system achieved an MTR of 4.7226 and Precision@1 of 98.34%, confirming high predictive reliability. Fairness behavior, examined via counterfactual influence distributions, yielded a genre-level coefficient of variation (CV) of 0.7232, indicating heterogeneous but interpretable fairness sensitivity across content categories. Genre removals showed that *Drama*, *Action*, and *Comedy* exert the strongest influence on group utilities. The average recommendation change rate was 15.98%, implying an overall stability rate of 84.02%. Tag-level analyses (CV $\approx$ 0.53) corroborated the macro-level findings at finer granularity, with consistent influence rankings across 100 user-generated tags. These findings are limited to the evaluated MovieLens dataset using synthetically sampled user groups and should not be interpreted as evidence that the proposed framework outperforms alternative group recommendation methods. Taken together, these results suggest that NBS can yield fair, accurate, and robust group recommendations under the evaluated MovieLens setting. Future work will include comparisons with standard group recommendation baselines, additional fairness measures, and statistical significance testing.

References

1. Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B.: The connection between popularity bias, calibration, and fairness in recommendation. ACM RecSys (2020)
2. Benkessirat, S., Boustia, N., Nachida, R.: A new collaborative filtering approach based on game theory for recommendation systems. Journal of Web Engineering **20**(2), 303–326 (2021)

3. Borges, R., Stefanidis, K.: F2VAE: a framework for mitigating user unfairness in recommendation systems. In: ACM/SIGAPP SAC. pp. 1391–1398 (2022)
4. Garcia, I., Sebastia, L.: A negotiation framework for heterogeneous group recommendation. Expert Systems with Applications **41**(4, Part 1), 1245–1261 (2014)
5. Giannopoulos, G., Papastefanatos, G., Sacharidis, D., Stefanidis, K.: Interactivity, fairness and explanations in recommendations. In: UMAP. pp. 157–161 (2021)
6. He, M., An, B., Wang, J., Wen, H.: Cetd: Counterfactual explanations by considering temporal dependencies in sequential recommendation. Applied Sciences (2023)
7. Le, G.H.T., Nguyen-Tan, K., Huynh, H.X.: Energy-based fairness recommendation: A novel framework for fairness treatment in recommender systems. ACIE (2025)
8. Lenzi, E., Stefanidis, K.: Adapt: Fairness & diversity for sequential group recommendations. Information Systems **133**, 102572 (2025)
9. Li, Y., Chen, H., Fu, Z., Ge, Y., Zhang, Y.: User-oriented fairness in recommendation. The Web Conference 2021 (2021)
10. Nash, J.F.: The bargaining problem. Econometrica **18**(2), 155–162 (1950)
11. Papamitsiou, Z., Economides, A.A.: Recommendation of educational resources to groups: A game-theoretic approach. In: IEEE Global Engineering Education Conference. pp. 760–767 (2018)
12. Pitoura, E., Stefanidis, K., Koutrika, G.: Fairness in rankings and recommenders: Models, methods and research directions. In: ICDE. pp. 2358–2361. IEEE (2021)
13. Polyzou, A., Kalantzi, M., Karypis, G.: Faireo: User group fairness for equality of opportunity in course recommendation. ArXiv **abs/2109.05931** (2021)
14. Ranjbar, N., Momtazi, S., Homayounpour, M.: Explaining recommendation system using counterfactual textual explanations. Mach. Learn. **113**, 1989–2012 (2023)
15. Shen, Y., Ma, X., Bao, Y., Xu, Z., Zhan, J.: Strategic manipulation behavior analysis for group decision-making based on nash bargaining game and regret theory. IEEE Trans. on Systems, Man, and Cybernetics: Systems **55**(10), 6814–6828 (2025)
16. Stratigi, M., Bikakis, N., Stefanidis, K.: Counterfactual explanations for group recommendations. In: DOLAP. pp. 22–30. CEUR Workshop Proceedings (2025)
17. Tan, J., Xu, S., Ge, Y., Li, Y., Chen, X., Zhang, Y.: Counterfactual explainable recommendation. CIKM (2021)
18. Tansuchat, R., Kosheleva, O.: How to make recommendation systems fair: an adequate utility-based approach. Asian Journal of Economics and Banking (2022)
19. Wang, X.J., Li, Q., Yu, D., Li, Q., Xu, G.: Counterfactual explanation for fairness in recommendation. ACM Transactions on Information Systems **42**, 1 – 30 (2023)
20. Wei, Y.W., Qu, X., Wang, X., Ma, Y., Nie, L., seng Chua, T.: Rule-guided counterfactual explainable recommendation. IEEE TKDE **36**, 2179–2190 (2024)
21. Zhang, L., Zhou, R., Jiang, H., Wang, H., Zhang, Y.: Item group recommendation: A method based on game theory. In: WWW Companion (2017)
22. Zhao, H., Liu, Q., Ge, Y., Kong, R., Chen, E.: Group preference aggregation: A nash equilibrium approach. In: ICDM. pp. 679–688 (2016)