

T.C.
AKDENİZ ÜNİVERSİTESİ



**RASTGELE BOZMA YAKLAŞIMLARININ ÖNERİ ALGORİTMALARININ
POPÜLERLİK YANLILIĞINA ETKİSİNİN İNCELENMESİ VE SAHTE OY
ENJEKSİYONUNA DAYALI YENİLİKÇİ ÇÖZÜMLER GELİŞTİRİLMESİ**

Mert GÜLSOY

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DOKTORA TEZİ

MART 2024

ANTALYA

T.C.
AKDENİZ ÜNİVERSİTESİ



**RASTGELE BOZMA YAKLAŞIMLARININ ÖNERİ ALGORİTMALARININ
POPÜLERLİK YANLILIĞINA ETKİSİNİN İNCELENMESİ VE SAHTE OY
ENJEKSİYONUNA DAYALI YENİLİKÇİ ÇÖZÜMLER GELİŞTİRİLMESİ**

Mert GÜLSOY

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DOKTORA TEZİ

MART 2024

ANTALYA

T.C.
AKDENİZ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

RASTGELE BOZMA YAKLAŞIMLARININ ÖNERİ ALGORİTMALARININ
POPÜLERLİK YANLILIĞINA ETKİSİNİN İNCELENMESİ VE SAHTE OY
ENJEKSİYONUNA DAYALI YENİLİKÇİ ÇÖZÜMLER GELİŞTİRİLMESİ

Mert GÜLSOY

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DOKTORA TEZİ

Bu tez çalışması TÜBİTAK tarafından 122E040 nolu proje kapsamında
desteklenmiştir.

MART 2024

T.C.
AKDENİZ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

RASTGELE BOZMA YAKLAŞIMLARININ ÖNERİ ALGORİTMALARININ
POPÜLERLİK YANLILIĞINA ETKİSİNİN İNCELENMESİ VE SAHTE OY
ENJEKSİYONUNA DAYALI YENİLİKÇİ ÇÖZÜMLER GELİŞTİRİLMESİ

Mert GÜLSOY

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DOKTORA TEZİ

Bu tez 08/05/2024 tarihinde jüri tarafından Oybirliği ile kabul edilmiştir.

Doç. Dr. Alper BİLGE (Danışman)

Doç. Dr. Alper Kürşat UYSAL

Doç. Dr. Alper ÖZCAN

Dr. Öğr. Üyesi Murat AK

Dr. Öğr. Üyesi Özge ÖZTİMUR KARADAĞ

ÖZET

RASTGELE BOZMA YAKLAŞIMLARININ ÖNERİ ALGORİTMALARININ POPÜLERLİK YANLILIĞINA ETKİSİNİN İNCELENMESİ VE SAHTE OY ENJEKSİYONUNA DAYALI YENİLİKÇİ ÇÖZÜMLER GELİŞTİRİLMESİ

Mert GÜLSOY

Doktora Tezi, Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Alper BİLGE

İkinci Danışman: Doç. Dr. Emre YALÇIN

Mart 2024, 133 sayfa

Bu çalışma, gizlilik korumalı işbirlikçi filtreleme sistemleri ve popülerlik önyargısının azaltılması üzerine yoğunlaşmıştır. İşbirlikçi filtreleme, kullanıcıların geçmiş etkileşimlerinden yararlanarak kişiselleştirilmiş öneriler sunan önemli bir öneri sistemi yöntemidir. Ancak, bu sistemler, kullanıcı verilerinin gizliliğini tehlikeye atabilir ve popüler öğelerin yoğun bir şekilde temsil edilmesi gibi durumlara neden olabilir. Gizlilik korumalı işbirlikçi filtreleme, kullanıcı verilerinin gizliliğini korurken etkili öneriler sağlama amacı güder. Bu çalışma, gizliliği koruyan işbirlikçi filtreleme yöntemleri arasında, rastgele bozulma ve obfuskasyon tekniklerini kullanarak kullanıcı tercihlerini anonimleştiren yöntemleri ele alır. Popülerlik önyargısı, popüler öğelerin aşırı önerilmesi sonucu öneri çeşitliliğinin azalmasına neden olan bir sorundur. Popülerlikten arındırma (popularity debiasing), bu önyargıyı azaltmayı ve öneri listelerindeki çeşitliliği artırmayı hedefler. *EquiRate* yöntemi, popülerlik önyargısını azaltmaya yönelik yenilikçi bir yaklaşım sunar. Bu yöntem, öneri listelerindeki çeşitliliği ve adilliliği artırarak popüler olmayan öğelere daha fazla şans verir. *FusionIndex* metriği, öneri sistemlerinin başarısını hem doğruluk hem de çeşitlilik açısından değerlendiren kapsamlı bir ölçümdür. Yapılan deneyler, *EquiRate*'ın ve özellikle *FusionIndex* metriği kullanılarak değerlendirildiğinde, popülerlikten arındırma yöntemlerine kıyasla önemli iyileştirmeler sunduğunu göstermiştir. Bu sonuçlar, öneri sistemlerinde popülerlik önyargısının azaltılmasının yanı sıra, gizlilik korumalı işbirlikçi filtreleme yöntemlerinin etkinliğini de ortaya koymaktadır. Bu tez, öneri sistemlerinde gizlilik koruması ve popülerlik önyargısının azaltılması konularında önemli

katkılar sunar. Gizlilik korumalı işbirlikçi filtreleme ve popülerlikten arındırma yaklaşımlarının birleştirilmesi, hem kullanıcı gizliliğini korurken hem de öneri çeşitliliğini ve adilliliğini artıran etkili bir strateji olarak öne çıkmaktadır. Bu çalışma, öneri sistemlerinin gelecekteki gelişimine yönelik değerli içgörüler ve yönlendirmeler sağlar.

ANAHTAR KELİMELER: Adalet, Gizlilik Koruma, Ortak Filtreleme, Ön-işlem, Popülerlik Yanlılığı, Sentetik Oy Enjeksiyonu.

JÜRİ: Doç. Dr. Alper BİLGE

Doç. Dr. Alper Kürşat UYSAL

Doç. Dr. Alper ÖZCAN

Dr. Öğr. Üyesi Murat AK

Dr. Öğr. Üyesi Özge ÖZTİMUR KARADAĞ

ABSTRACT

ANALYZING EFFECTS OF RANDOM PERTURBATION APPROACHES ON THE POPULARITY BIAS ISSUE OF RECOMMENDATION ALGORITHMS AND DEVELOPING NOVEL FAKE RATING INJECTION-BASED SOLUTIONS

Mert GÜLSOY

PhD Thesis in COMPUTER SCIENCES

Supervisor: Assoc. Prof. Dr. Alper BİLGE

Second Supervisor: Assoc. Prof. Dr. Emre YALÇIN

March 2024, 133 pages

This thesis explores the advancement of recommender systems, with a focus on addressing popularity bias through privacy-preserved collaborative filtering techniques. Recommender systems are pivotal in filtering vast information spaces, guiding users towards items of interest. However, these systems often suffer from popularity bias, where popular items are disproportionately recommended, overshadowing less known items. This work introduces two novel approaches: a robust method for privacy-preserved collaborative filtering and the *EquiRate* algorithm for mitigating popularity bias. The privacy-preserved collaborative filtering technique employs randomized perturbation and obfuscation to safeguard user privacy while maintaining recommendation quality. This method not only enhances user trust by protecting sensitive information but also contributes to the accuracy and reliability of the recommendations. On the other hand, the *EquiRate* method specifically addresses the challenge of popularity bias. By integrating the *FusionIndex* metric, which assesses both recommendation accuracy and diversity, *EquiRate* efficiently balances the representation of popular and *niche* items, promoting a more diverse and equitable item exposure. Experimental evaluations on benchmark datasets reveal that these methods significantly improve recommendation diversity without compromising accuracy. The robust privacy-preserved collaborative filtering demonstrates resilience against various privacy attacks, ensuring effective recommendation under stringent privacy constraints. Meanwhile, *EquiRate* outperforms existing popularity-debiasing methods across multiple datasets, as evidenced by its superior *FusionIndex* scores. These outcomes

highlight the potential of integrating privacy preservation and bias mitigation strategies to enhance the overall effectiveness of recommender systems. In conclusion, this thesis contributes to the recommender system literature by presenting innovative solutions for two pressing issues: privacy preservation and popularity bias. The proposed methods not only advance the state-of-the-art in collaborative filtering but also pave the way for creating more balanced, accurate, and privacy-conscious recommendation platforms.

KEYWORDS: collaborative filtering, fairness, popularity bias, pre-processing, privacy-preserving, synthetic rating injection.

COMMITTEE: Assoc. Prof. Dr. Alper BİLGE

Assoc. Prof. Dr. Alper Kürşat UYSAL

Assoc. Prof. Dr. Alper ÖZCAN

Asst. Prof. Dr. Murat AK

Asst. Prof. Dr. Özge ÖZTİMUR KARADAĞ

ÖNSÖZ

Information retrieval araştırma alanı altında yer alan, ancak zaman içerisinde gelişerek kendi başına önemli bir ihtisas olan "*Öneri Sistemleri*" konusu ile ilk tanıştığımda beni fazlasıyla heyecanlandırdı. Bunun nedeni, sadece hesaplama bilimlerindeki yapay zeka tekniklerinin kullanılması değildi. Aynı zamanda bu araştırma alanı; tüketim davranışları, pazarlama ve ticaret gibi insanı içine alan iktisadi konuları da dolaylı olarak içermesiydi. Bu durum nedeniyle öneri sistemleri bende kişisel olarak derin bir ilgi uyandırmıştır.

İlgili araştırma konusu dünyanın en ilgi çekici konusu olsa da, araştırma danışmanının da rolü en az konu kadar önemlidir. Değerli hocam Doç. Dr. Alper BİLGE'nin, sadece öneri sistemleri konusu ile beni tanıştırmaması değil, konu ile ilgili bilgi aktarımı, sabrı, öngörüsü, yönlendirmesi gibi bir hocanın sağlaması gereken her yaklaşımı sunmuş ve süreç içerisinde destek olmuştur. Bunun için kendisine minnettar durumdayım.

Değerli hocam Doç. Dr. Alper BİLGE'nin sayesinde yaşadığım bir diğer dönüm noktası ise, Doç. Dr. Emre YALÇIN ile tanıştırmasıdır. Doç. Dr. Emre YALÇIN; gerek araştırma sürecinde, gerek makale çalışmasında bana sadece öğrenci olarak değil, aynı zamanda bir meslektaş olarak yaklaşmış, ayırdığı zaman ve gösterdiği sabır ile "*Nasıl öğrenci yetiştirilir?*" sorusunun ötesinde "*Araştırmacı nasıl yetiştirilir?*" sorusunun bir cevabı olmuştur. Umarım bundan sonra yürüteceğim araştırmalar da bu sorunun en güzel cevabı olacaktır.

Doktora sürecinde başarı ile tamamlanan dersler, yeterlilik sınavı, tez döneminde yapılan araştırmalar ve yürütülen deneylerin ciddi emek ve zaman gerektirmiştir. Bu çalışmaları yürütebilmek için gereken bu zaman ve emeği kendilerinden esirgemek durumunda kaldığım eşim Dr. Öğr. Üyesi Hatice GÜLSOY ve biricik kızlarım İpek GÜLSOY ile İlay GÜLSOY'a beni anlayıp destek verdikleri için çok teşekkür ediyorum.

TABLE OF CONTENTS

ÖZET	i
ABSTRACT	iii
ÖNSÖZ	v
AKADEMİK BEYAN	viii
SYMBOLS AND ABBREVIATIONS	ix
LIST OF FIGURES	x
LIST OF TABLES	xiii
1. INTRODUCTION	1
1.1. Problem Statement and Motivation	2
1.2. Purpose and Contributions	7
1.3. Thesis Organisation	11
2. PRELIMINARIES	13
2.1. Recommender Systems	13
2.2. Collaborative Filtering	23
2.3. Privacy Concerns in Recommender Systems	31
2.3.1. Methodologies for Enhancing Privacy	33
2.3.2. Randomized Perturbation and Obfuscation	35
2.3.3. The Long-Tail Model	37
2.4. Popularity Bias Problem and Debiasing Methods	39
2.5. Datasets	51
2.6. Evaluation Methods	54
2.6.1. Accuracy Metrics	55
2.6.2. Beyond Accuracy Metrics	56
3. LITERATURE REVIEW	60
3.1. Privacy-Preserving Approaches for CF Recommenders	60
3.2. Popularity Bias Issue of CF Recommenders	64
3.3. Popularity Debiasing Solutions	67
4. MATERIALS AND METHODS	73
4.1. Constructing User Personas Based on Varying Privacy Protection Requirements	73
4.2. Experimental Setup	77
4.2.1. Evaluation Metrics	78

4.2.2. Experimentation Methodology	78
4.2.3. The Benchmark Treatment Methods for Popularity Bias Problem	80
4.3. Analyzing Robustness of Privacy-Preserving Collaborative Recommenders Against Popularity Bias Issue	81
4.4. Analyzing How Popularity-Debiasing Approaches Differently Perform for Privacy Level-based User Personas	91
4.5. The <i>EquiRate</i> : The Proposed Balanced Rating Injection Strategy for The Popularity-Debiasing Problem	94
4.6. Experimental Studies	99
4.6.1. Parameter Setup	99
4.6.2. Evaluation Metrics	101
4.6.3. The <i>FusionIndex</i> : A Highly Useful Holistic Evaluation Metric	102
4.6.4. Experiment Results	103
4.6.5. In-depth Analysis of The Proposed <i>EquiRate</i> Method	103
4.6.6. Comparison of The <i>EquiRate</i> With Existing Popularity-Debiasing Methods	107
5. RESULTS AND DISCUSSION	116
6. CONCLUDING REMARKS	119
7. REFERENCES	122

CURRICULUM VITAE

AKADEMİK BEYAN

Doktora Tezi olarak sunduğum “Rastgele Bozma Yaklaşımlarının Öneri Algoritmalarının Popülerlik Yanlılığına Etkisinin İncelenmesi ve Sahte Oy Enjeksiyonuna Dayalı Yenilikçi Çözümler Geliştirilmesi” adlı bu çalışmanın, akademik kurallar ve etik değerlere uygun olarak bulunduğunu belirtir, bu tez çalışmasında bana ait olmayan tüm bilgilerin kaynağını gösterdiğimi beyan ederim.

18/03/2024

Mert GÜLSOY

İmza



SYMBOLS AND ABBREVIATIONS

Abbreviations:

APLT	: Average Percentage of Long Tail Items
APRI	: Average percentage of recommended items
CBF	: Content-Based Filtering
CF	: Collaborative Filtering
GDPR	: General Data Protection Regulation
GRV	: Generalized Rating Value
HIFA	: Head Item Focused Adjustment
HPF	: Hierarchical Poisson Factorization
IA	: Item Average
IBPR	: Indexable Bayesian Personalized Ranking
KBF	: Knowledge-Based Filtering
LTC	: Long Tail Coverage
MF	: Matrix Factorization
MMMF	: Maximum Margin Matrix Factorization
MP	: Most Popular
n DCG	: Normalized Discounted Cumulative Gain
NEUMF	: Neural Matrix Factorization
OPA	: Overall Popularity Adjustment
PPCF	: Privacy Preserved Collaborative Filtering
PRV	: Predictive Rating Value
RO	: Randomized Obfuscation
RP	: Randomized Perturbation
RS	: Recommender Systems
RVG	: Randomized Value Generation
SKM	: Spherical k -means
VAECF	: Variational Autoencoder for Collaborative Filtering
WBPR	: Weighted Bayesian Personalized Ranking
WMF	: Weighted Matrix Factorization

LIST OF FIGURES

Figure 1.1	The distribution of user preferences exhibiting a <i>long-tail</i> pattern in the data collections of (a) MovieLens-1M (movies), (b) Do- urban Book (books), and (c) Yelp (local businesses).	5
Figure 2.2	Main approaches in recommender systems	22
Figure 2.3	Personalization Benefits and Privacy Risks Trade-off	33
Figure 2.4	The Long-Tail Model	38
Figure 4.5	Randomized perturbation procedure for data concealing	74
Figure 4.6	Histograms depicting the estimated probabilities for three perso- nas within the (a) ML, (b) DB, and (c) Yelp datasets.	77
Figure 4.7	The <i>APRI</i> outcomes derived from the algorithms applied across user profiles categorized by four distinct levels of privacy within the (a) ML, (b) DB, and (c) Yelp datasets.	82
Figure 4.8	The <i>APLT</i> outcomes derived from the algorithms applied across user profiles categorized by four distinct levels of privacy within the (a) ML, (b) DB, and (c) Yelp datasets.	83
Figure 4.9	The <i>Entropy</i> outcomes derived from the algorithms applied ac- ross user profiles categorized by four distinct levels of privacy within the (a) ML, (b) DB, and (c) Yelp datasets.	85
Figure 4.10	The <i>Novelty</i> outcomes derived from the algorithms applied ac- ross user profiles categorized by four distinct levels of privacy within the (a) ML, (b) DB, and (c) Yelp datasets.	86
Figure 4.11	The <i>LTC</i> outcomes derived from the algorithms applied across user profiles categorized by four distinct levels of privacy within the (a) ML, (b) DB, and (c) Yelp datasets.	88
Figure 4.12	The derived <i>Precision</i> and <i>Recall</i> scores for the deployed algo- rithms across four distinct privacy level-based user profiles are detailed. <i>Precision</i> metrics are delineated in panels (a), (c), and (e) corresponding to the ML, DB, and Yelp datasets, respectively, whereas <i>Recall</i> metrics are depicted in panels (b), (d), and (f).	89

Figure 4.13	The derived <i>F1-score</i> and <i>nDCG</i> scores for the deployed algorithms across four distinct privacy level-based user profiles are detailed. <i>F1-score</i> metrics are delineated in panels (a), (c), and (e) corresponding to the ML, DB, and Yelp datasets, respectively, whereas <i>nDCG</i> metrics are depicted in panels (b), (d), and (f).	90
Figure 4.14	The enhancements or declines in (a) <i>APRI</i> , (b) <i>APLT</i> , (c) <i>Entropy</i> , (d) <i>Novelty</i> , and (e) <i>LTC</i> metrics observed upon the application of <i>Aug</i> , <i>Mul</i> , and <i>xQd</i> debiasing techniques for user personas classified by different privacy levels within the Yelp dataset.	92
Figure 4.15	The enhancements or declines in (a) <i>Precision</i> , (b) <i>Recall</i> , (c) <i>F1-score</i> , and (d) <i>nDCG</i> metrics observed upon the application of <i>Aug</i> , <i>Mul</i> , and <i>xQd</i> debiasing techniques for user personas classified by different privacy levels within the Yelp dataset.	93
Figure 4.16	Procedures of the proposed the <i>EquiRate</i> method	98
Figure 4.17	General structure of the proposed <i>EquiRate</i> approach	99
Figure 4.18	All parameter adjustments tested for the <i>EquiRate</i> are documented. The green dots symbolically illustrate each variant of the <i>EquiRate</i> .	101
Figure 4.19	The enhancement in the <i>FusionIndex</i> percentage through various implementations of the <i>EquiRate</i> methodology in the MLM dataset is documented. In this context, the different versions of <i>EquiRate</i> are identified based on the specific SRI, α , and SRG strategies that were employed.	104
Figure 4.20	The enhancement in the <i>FusionIndex</i> percentage through various implementations of the <i>EquiRate</i> methodology in the DB dataset is documented. In this context, the different versions of <i>EquiRate</i> are identified based on the specific SRI, α , and SRG strategies that were employed.	105

Figure 4.21	The enhancement in the <i>FusionIndex</i> percentage through various implementations of the <i>EquiRate</i> methodology in the Yelp dataset is documented. In this context, the different versions of <i>EquiRate</i> are identified based on the specific SRI, α , and SRG strategies that were employed.	106
Figure 4.22	The enhancement in the <i>FusionIndex</i> scores following the application of popularity-debiasing techniques for the MLM dataset.	108
Figure 4.23	The enhancement in the <i>FusionIndex</i> scores following the application of popularity-debiasing techniques for the DB dataset.	112
Figure 4.24	The enhancement in the <i>FusionIndex</i> scores following the application of popularity-debiasing techniques for the Yelp dataset.	115

LIST OF TABLES

Table 2.1	Example of a User-Item Matrix in Recommendation Systems . . .	14
Table 2.2	Detailed information about utilized recommendation algorithms .	28
Table 2.3	Characteristics of the MovieLens-1M (ML), Douban Book (DB), and Yelp datasets	54
Table 4.4	Parameters for the devised privacy-centric user profiles	75
Table 4.5	Privacy levels ($\Pi(P U)$) for different personas	78
Table 4.6	The assessment of the <i>EquiRate</i> variants in comparison with vari- ous popularity-debiasing techniques based on <i>nDCG</i> , <i>APLT</i> , <i>No-</i> <i>velty</i> , <i>LTC</i> , and <i>Entropy</i> metrics specifically for the MLM dataset.	109
Table 4.7	The assessment of the <i>EquiRate</i> variants in comparison with vari- ous popularity-debiasing techniques based on <i>nDCG</i> , <i>APLT</i> , <i>No-</i> <i>velty</i> , <i>LTC</i> , and <i>Entropy</i> metrics specifically for the DB dataset. . .	111
Table 4.8	The assessment of the <i>EquiRate</i> variants in comparison with vari- ous popularity-debiasing techniques based on <i>nDCG</i> , <i>APLT</i> , <i>No-</i> <i>velty</i> , <i>LTC</i> , and <i>Entropy</i> metrics specifically for the Yelp dataset. .	114

1. INTRODUCTION

Recommender systems constitute a significant technological breakthrough in the realm of artificial intelligence, fundamentally altering the way in which users engage with digital content and services. At their essence, these systems aim to identify and recommend items such as movies, books, or products that a user might find appealing, leveraging various types of data (Wang et al. 2023). Their deployment across a wide array of fields, from e-commerce to entertainment and social media, has markedly improved the user experience and interaction (Li et al. 2018). Through the analysis of extensive datasets encompassing user preferences, behaviors, and characteristics of items, recommender systems offer personalized content recommendations, establishing their indispensability in the modern digital ecosystem. This innovation not only elevates user satisfaction but also enhances business outcomes by enabling the delivery of more accurately targeted content and advertisements. The dynamic and expanding domain of recommender systems is in a state of continuous innovation, adopting advanced algorithms that range from collaborative filtering to deep learning techniques. This evolution exemplifies the ongoing quest for recommendation approaches that are more precise, aware of the context, and centered around the user (Batmaz et al. 2019).

With the emergence of online capabilities via the Internet, various digital platforms have been created to assist individuals in numerous areas, such as engaging in daily tasks (e.g., shopping, listening to music, watching films) or communicating with others through social media. Yet, the sheer volume of information on these platforms has escalated to what's known as the *information overload* challenge in recent times, complicating users' ability to navigate through it (Jannach and Zanker 2022). To address this issue, recommender systems (RSs) have been deployed as artificial intelligence-based tools that aim to match users' needs and preferences by recommending relevant items or services (Ko et al. 2022). These systems offer significant advantages to the primary stakeholders of these platforms, namely *users* and *providers of goods or services*. Consequently, leading

digital services across various sectors, including Netflix¹, Twitter², Spotify³, YouTube⁴, and Booking.com⁵, have incorporated these recommender tools into their frameworks, continually striving to enhance them to make more accurate and personalized recommendations to users.

1.1. Problem Statement and Motivation

Despite their widespread utilization and achievements, recommender systems are confronted with several substantial challenges that may hinder their efficiency (Ricci et al. 2015). *Data sparsity* stands out as a primary concern; even in the presence of extensive datasets, the quantity of user-item interactions frequently falls short, posing difficulties in formulating precise recommendations. The *cold-start* issue, attributing to insufficient interaction data for new users or items, presents another significant barrier. The robustness of these systems is crucial, necessitating resilience against data manipulations or irregularities to safeguard the credibility of their recommendations. Moreover, prevalent biases, such as those related to users or items, have the potential to distort recommendations in favor of certain products or user groups, thus restricting the diversity of the recommendations (Chen et al. 2023).

Collaborative filtering (CF) methods stand out as the most effective strategies for generating personalized recommendations (Su and Khoshgoftaar 2009), yet there are also specialized approaches for certain scenarios, such as *content-* (Shambour et al. 2022) or *knowledge-based filtering* (Agarwal et al. 2022). Through the analysis of a matrix that maps user-item preferences, collaborative recommendation systems primarily leverage similarities among users or items to carry out the recommendation process. This is done either by (i) calculating prediction scores for items that users have not yet encountered or (ii) creating sorted lists (i.e., top- N lists) that contain items recommended for users (Bobadilla et al. 2013). In the past two decades, numerous algorithms employing this methodology have been introduced, leading to a comprehensive body of literature on

¹<https://netflix.com/>

²<https://twitter.com/>

³<https://spotify.com/>

⁴<https://youtube.com/>

⁵<https://booking.com/>

their principal drawbacks (Koren et al. 2022). Although initial research predominantly focused on addressing issues related to *scalability*, *precision*, or the *cold-start* problem, the tendency of these systems to recommend a narrow set of popular items while neglecting a wider array has emerged as a significant challenge known as the ***popularity-bias*** issue, gaining critical attention in the realm of recommendation systems in recent times (Abdollahpouri et al. 2020; Elahi et al. 2021).

The spread of such undesirable bias by CF recommendation systems can pose significant issues for both users and service providers in practical scenarios. On one side, users may not access high-quality recommendations that cater to aspects beyond accuracy, like *diversity*, *serendipity*, and *novelty*, leading to a notable decline in their contentment with the service (Yalcin and Bilge 2022). On the flip side, the distribution of items in the recommendations does not represent the entire catalogue equally, resulting in an imbalanced market competition. Even worse, if the system harbors this flaw, it could become a target for manipulative schemes by nefarious actors aiming to promote their products or undermine competitors' offerings for financial gain (Si and Li 2020). Moreover, in specific domains such as social media, this vulnerability could be exploited to sway users into ideological bubbles (Cinelli et al. 2021). Due to these factors, numerous studies in the field of recommendation systems have thoroughly examined the negative impacts of *popularity bias* and endeavored to devise a range of mitigation strategies, known as ***popularity-debiasing*** techniques, aiming to achieve more equitable recommendation lists in terms of item popularity (Abdollahpouri et al. 2019; Yalcin and Bilge 2021; Boratto et al. 2021). Although various *popularity-debiasing* approaches are available, *post-processing* techniques, which involve adjusting recommendation lists by demoting popular items, are particularly prevalent because they do not rely on the underlying recommendation model and can be seamlessly integrated into the output of any recommendation system.

Among the various challenges faced by recommender systems, the issue of *popularity bias* stands out as particularly detrimental (Abdollahpouri et al. 2019; Yalcin and Bilge 2021; Elahi et al. 2021). This bias leads to a disproportionate preference for popular items by recommender systems, resulting in recommendations that are narrowly focused. Such a tendency not only diminishes the visibility of lesser-known or new items but also restricts the diversity and innovation within the recommendations offered. The adverse

effects of *popularity bias* are not limited to reduced exposure for certain items; it also results in a homogenization of the user experience, failing to adequately reflect the diverse preferences of users. This uniformity can further fuel a feedback loop, making popular items even more prevalent and sidelining *niche* content. Consequently, the overall user experience suffers, with a significant reduction in the opportunities for discovery and the provision of personalized content. Therefore, addressing *popularity bias* is a critical goal in the development of recommender systems, with the aim of fostering a recommendation environment that is more equitable, varied, and centered around the user (Abdollahpouri et al. 2021; Boratto et al. 2021; Yalcin and Bilge 2022).

The prevailing consensus on the emergence of bias against popular items in recommendations points to the imbalances found in the rating distribution of the original preference data used to train the recommendation algorithm (Abdollahpouri et al. 2018). More specifically, a limited number of *popular items* tend to receive the majority of ratings in a dataset. In contrast, the vast majority of items receive attention from only a small group of users, leading to a distribution of user preferences that spreads out in a *long-tail* distribution across items, as depicted in Fig. 1.1 for three benchmark datasets from different application areas.

The fundamental cause of *popularity bias* in recommender systems is rooted in the unequal distribution of user ratings. Typically, a minority of items amass a disproportionately large number of interactions and ratings from users, leaving the majority with scant attention. This pattern mirrors actual user behavior, where items that are already popular tend to receive more focus and feedback. However, this imbalance, when inputted into recommender systems, exacerbates the *popularity bias*. The algorithms, designed to prioritize accuracy and relevance, inherently prefer items with more interactions, under the assumption that these are more likely to appeal to a wider audience. Fig. 1.1 demonstrates the disparities in rating distributions among three different datasets, each representing distinct real-world application fields (Gulsoy et al. 2023).

The undue emphasis on popular items initiates a self-perpetuating cycle, where the frequency of recommendations and interactions for a *popular item* increases its chances of future recommendations. This dynamic effectively sidelines less popular or new items, potentially highly relevant to specific user groups, due to their lower levels of ini-

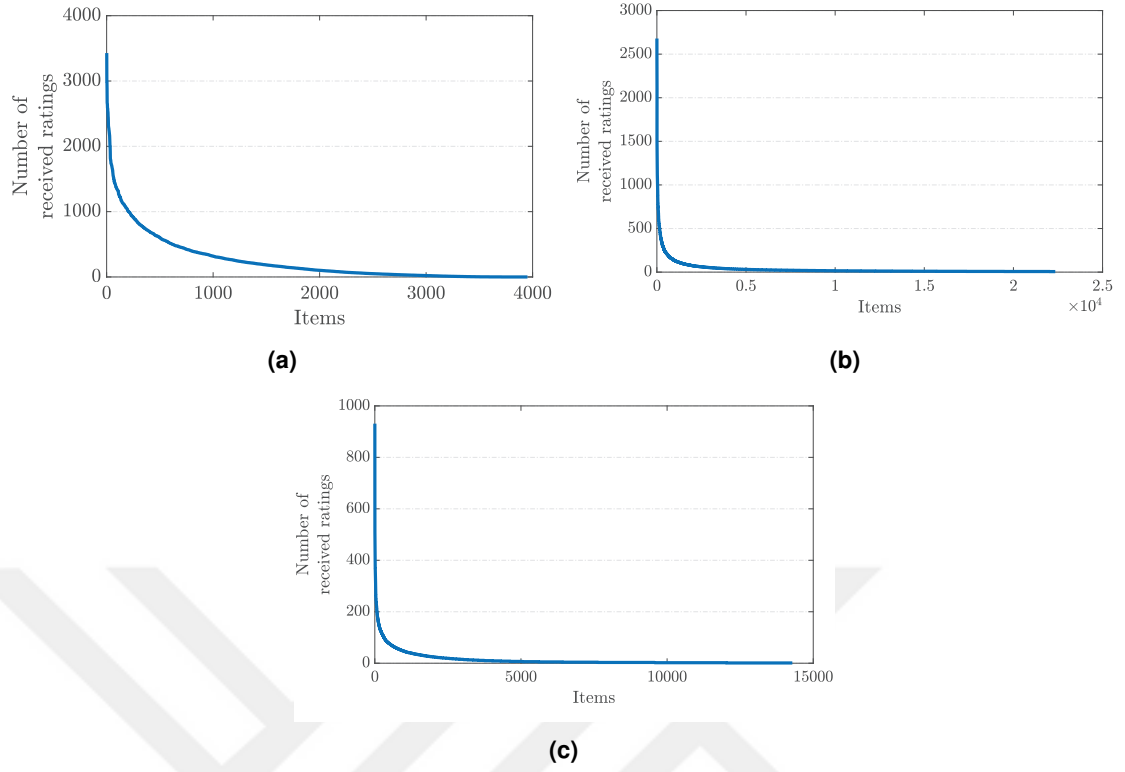


Figure 1.1. The distribution of user preferences exhibiting a *long-tail* pattern in the data collections of (a) MovieLens-1M (movies), (b) Douban Book (books), and (c) Yelp (local businesses).

tial engagement. This biased distribution of user interactions results in a recommendation landscape that is unduly narrow, favoring popular items to a significant degree (Abdollahpouri and Burke 2019). The challenge lies in the development of recommender systems capable of identifying and mitigating this imbalance, to ensure the broader spectrum of items, especially the *long-tail* of less popular items, is adequately represented. By tackling the fundamental issue of skewed user rating distributions, recommender systems can evolve to provide a richer, more inclusive, and customized recommendation experience, overcoming the limitations imposed by *popularity bias* (Abdollahpouri et al. 2017; Abdollahpouri et al. 2019; Borges and Stefanidis 2021).

The domain of counteracting *popularity bias* within recommender systems is delineated into three principal methodologies: *post-processing*, *in-processing*, and *pre-processing*. Each strategy offers distinct sets of merits and demerits (Boratto et al. 2021; Yalcin and Bilge 2021). *Post-processing* methods are deployed after the generation of recommendati-

ons, mainly focusing on the re-ranking of items to foster a more equitable item distribution (Abdollahpouri et al. 2019; Gupta et al. 2023). Despite its intent to balance out item visibility, it tends to be somewhat superficial by not addressing the inherent bias within the algorithm, merely modifying its final recommendations. Conversely, *in-processing* techniques incorporate bias mitigation directly into the algorithmic mechanism (Kamishima et al. 2014; Hou et al. 2018a), potentially adjusting the algorithm to diminish the favoritism towards *popular items* or to enhance diversity actively. This method allows for a more intrinsic approach to bias reduction during the model’s training process but may necessitate comprehensive modifications to the algorithm and could lead to increased computational demands.

Pre-processing techniques, emphasized within the framework of this study, entail the adjustment of input data prior to its application within the recommendation model (Janach et al. 2015; Chen et al. 2018). The cardinal merit of employing this strategy lies in its capability to directly tackle the foundational cause of bias—the imbalanced distribution of user ratings—by either re-distributing or enriching the dataset. This endeavor has the potential to ensure a more balanced representation across the spectrum of items, thereby facilitating a fairer exposure for both widely recognized and less known items within the recommender system.

One significant challenge facing *collaborative filtering* (CF) recommendation systems is the difficulty in gathering users’ sensitive preference data, which can stem from various factors such as poor user interface designs and user apathy. This problem becomes even more acute in certain domains like social media, where individuals are reluctant to share their views on controversial topics due to fears of data breaches (Lu et al. 2022). The inability to collect accurate user profiles typically results in subpar recommendations, particularly in terms of prediction accuracy, since traditional CF systems rely on data comprising untrustworthy ratings (Yargic and Bilge 2019). Consequently, there’s a growing interest in *privacy-preserving collaborative filtering* (PPCF) methods that aim to deliver precise forecasts without compromising users’ privacy concerns (Ozturk and Polat 2015; Pramod 2022). While various PPCF strategies are available, the leading method in centralized server-based recommendation systems involves employing *random perturbation and obfuscation techniques* during the preference collection phase. These techniques

conceal (i) the list of items actually rated and (ii) the true ratings assigned to these items (Polat and Du 2003; Yargic and Bilge 2019). As perturbation involves adding random noise based on a specifically chosen probability distribution, it allows the central server to generate reliable and private recommendations. Thus, PPCF algorithms utilize a modified version of the original preference data, where the degree of perturbation is carefully calibrated to balance the competing objectives of maintaining accuracy and ensuring privacy.

Inserting random noise into user profiles effectively masks the list of items that have been rated by incorporating fictitious votes in some unrated slots, potentially causing significant shifts in how ratings are distributed among items. As a result, this manipulation can drastically change the observed level of popularity bias in recommendations, primarily because the core of this issue lies in the uneven spread of user preferences. Furthermore, the addition of synthetic ratings to user profiles changes the total count of ratings that each item receives. Consequently, one would anticipate notable variations in the effectiveness of methods designed to address popularity bias, particularly those employing *post-processing* techniques. These methods largely focus on adjusting items' positions in recommendation lists by penalizing them according to their popularity (i.e., the number of ratings they have received) (Abdollahpouri et al. 2019; Yalcin and Bilge 2021). Despite the critical importance of *privacy-preserving* recommendation systems in the current era of heightened data protection awareness, the dynamics of how they propagate *popularity bias* are not fully understood, and the efficacy of established methods for mitigating *popularity bias* within these systems remains unexplored.

1.2. Purpose and Contributions

In this thesis, aim to offer a thorough analysis on how *privacy-preserving* techniques based on *randomized perturbation* influence the spread of *popularity bias* in recommendation systems. Also examine three well-known methods for reducing *popularity bias* and assess their effectiveness under different levels of privacy protection for users. To this end, define four user personas based on their preferred level of privacy protection. The first persona, *Utterly Unconcerned (UU)*, denotes users who are indifferent to their privacy and seek the highest level of accuracy in the recommendations they receive. It's worth noting that *UU* also refers to the users of traditional *collaborative filtering* re-

commenders without privacy concerns. The second persona, known as *Barely Bothered (BB)*, is for users who are minimally concerned about privacy breaches and opt for low privacy settings in exchange for better accuracy. The third persona, *Pretty Pragmatists (PP)*, represents users who value their privacy but are also interested in receiving quality recommendations, thus they choose privacy settings that offer a compromise between *accuracy* and *privacy*. Lastly, the *Fiery Fundamentalists (FF)* persona is characterized by users who prioritize their privacy above all and are prepared to accept highly distorted recommendations in favor of protecting their privacy.

At the same time, in this thesis, present a pioneering pre-processing technique against *popularity bias*, dubbed *EquiRate*. This method is specifically designed to systematically integrate *synthetic ratings* for items on the lower end of the popularity spectrum, or the *tail items*, within the dataset. This methodology tackles two principal challenges: the first challenge involves identifying which *tail items* should receive *synthetic ratings* and determining the precise quantity of these ratings; the second challenge concerns formulating methods to accurately assign the values of these *synthetic ratings* in such a manner that does not alter the original rating distributions of the items.

To tackle the initial challenge, this approach utilizes a data-driven technique to gauge the imbalance present within the user-item interaction dataset. Based on this analysis, then determine the optimal quantity of *synthetic ratings* required to sufficiently represent *tail items*, thus enhancing their presence in the recommender system without saturating the overall dataset. Regarding the second challenge, calculate *synthetic ratings* with a focus on preserving the authenticity and coherence of the original rating vectors. Proposed algorithm is designed to carefully produce ratings that are consistent with the patterns and preferences indicated by actual user data. This meticulous process ensures that the introduced *synthetic ratings* naturally integrate with existing genuine ratings, maintaining the authentic user-item interaction dynamics. By effectively addressing these challenges, The proposed *EquiRate* methodology aspires to finely balance the augmentation of less popular items' representation with the preservation of the dataset's inherent structure. This strategy not only addresses the issue of *popularity bias* but also enhances the diversity and quality of the recommendations provided, fostering a more equitable and user-centric recommendation system.

The main research topic of this study is "*Developing New Methods to Reduce the Negative Effects of Popularity Bias with PPCF techniques in Recommender Systems*". However, in order to fully investigate this topic, it had to be divided into various phases. For this, the study was phased with four main research questions. to mention these research questions;

- **RQ-1:** How do randomized *perturbation-based privacy-preserving* approaches affect the level of *popularity bias* in *collaborative filtering* algorithms?
- **RQ-2:** Does a *randomized perturbation-based privacy-preserving collaborative filtering* approach mitigate the impact of *popularity bias*? Is it effective, robust, or does it exhibit resilience?
- **RQ-3:** Is it possible to develop a new *pre-processing* debiasing method inspired by the *synthetic rating injection* approach of *randomized perturbation* methods?
- **RQ-4:** Can be develop a more advanced evaluation metric that measures both the *accuracy* and the *beyond-accuracy* aspects of recommendations concurrently?

In this study, two different contributions were identified and investigated in relation to the RQ-1. These contributions are mainly planned within the scope of experiments to be carried out on profiles of users injected with different densities of fake ratings.

- Conduct an extensive experimental investigation to assess the extent of *popularity bias* dissemination in recommendation outputs across four distinct scenarios—*Utterly Unconcerned (UU)*, *Barely Bothered (BB)*, *Pretty Pragmatists (PP)*, and *Fiery Fundamentalists (FF)*—utilizing ten cutting-edge recommendation algorithms from various classes, tested on three benchmark datasets.
- Explore the variation in the negative impact of *popularity bias* on both the *accuracy* and the quality of recommendations that go *beyond accuracy*, for user personas categorized by different levels of privacy concerns.

In this study, one contribution was identified and investigated in relation to the RQ-2. This contribution is also planned within the scope of experiments to be carried out on

profiles created from users injected with fake ratings at different densities. However, in this phase, unlike the previous one, the effect levels were also measured.

- Evaluate three renowned methods for mitigating *popularity bias* to see if their efficacy is consistent across scenarios with varying degrees of privacy protection, such as *BB*, *PP*, and *FF*, compared to a scenario with no privacy protection (*UU*).

In order to answer RQ-3 whether a new debiasing method inspired by *privacy-preserving collaborative filtering* methods and synthetic rating injection approach can be developed, have identified three different contributions. The principal contributions of this study, which offers a groundbreaking method to mitigate *popularity bias* in recommender systems, are outlined as follows:

- ***Novel method for mitigating popularity bias:*** Introduce an innovative *pre-processing* strategy, namely *EquiRate*, that incorporates *synthetic ratings* for less popular (*tail*) items within the dataset. This technique addresses the challenge of *popularity bias* by improving the representation of these items, facilitating a more equitable recommendation mechanism.
- ***Strategic incorporation of synthetic ratings:*** This study introduces a methodical framework for the identification and selection of tail items for *synthetic rating injections*, along with the calculation of the ideal number and values of these ratings. This approach ensures that the authenticity of the original rating vectors is upheld, while also preserving the genuine patterns of user-item interactions.
- ***Promotion of recommendation diversity and equity:*** By adjusting the rating distribution to favor both popular and less popular items, proposed *EquiRate* technique significantly enhances the *diversity* and *fairness* of recommendations. This contributes to an enriched user experience by promoting the visibility of a wider selection of items, including those that are *niche* yet pertinent.

Within the scope of RQ-4, the following contribution was determined for the development of a new metric that handles *accuracy* and *beyond accuracy* metrics at the same time in order to measure the success of the research.

- ***The FusionIndex - A comprehensive evaluation metric for recommender systems:***

Introduce the *FusionIndex*, a novel metric designed to evaluate the quality of recommendation lists by effectively blending *accuracy* with other critical aspects such as *novelty*, *diversity*, *catalog coverage*, and *fairness*, utilizing a comprehensive combination of various evaluation metrics.

The scope of some of the contributions identified at the beginning of the study has been updated in the process to ensure that they fully cover the main topic and fully illuminate the research, and are aligned with the original purpose.

1.3. Thesis Organisation

In this thesis, there are eight main chapters in total. In the first chapter, the problem and motivation are mentioned together with the general introduction. In the rest of the chapter, the purposes required for the solution of the problem are identified. Then, contributions are derived from these purposes and related to the objectives.

In the second part of the thesis, general information is given. In this section, recommendation systems and general approaches of recommendation systems, Collaborative filtering and algorithms are mentioned so that readers can comprehend the research more easily. Due to the subject matter of the study, the privacy concern of users in recommender systems is explained and methods to reduce this sensitivity are discussed. In addition, popularity bias and *long-tail* concepts are detailed in this section. At the end of the chapter, datasets and evaluation metrics are explained in detail.

In the third chapter, the literature is reviewed, related works are identified and analyzed in search of potential answers to the research questions. Here, as in the general information chapter, collaborative filtering, privacy preserved collaborative filtering, popularity bias issue and debiasing approaches are discussed in depth.

The fourth chapter conclude that; Delves into the effectiveness of privacy-preserving measures against the challenge of popularity bias, followed by an examination of how different popularity bias mitigation approaches perform across user personas defined by their privacy preferences. It is noted that the source code associated with this research is made available to the public for purposes of reproducibility⁶.

⁶<https://doi.org/10.5281/zenodo.7789914>

Fifth chapter conclude that presents the results of the conducted experiments, comparing the performance of *EquiRate* against other debiasing methods and underlining its benefits. This section also elaborates on evaluation metrics, specifically emphasizing the newly proposed *FusionIndex* metric additionally.

In the sixth section, all the results of the experiments were collected and analyzed. These analyses were transformed into findings and then discussed. Afterwards, the results of the discussion are related to the contributions that were determined before.

In the seventh chapter, the results are discussed with a focus on the research questions and the study is concluded with a scientific output. It also includes notes for potential future studies.



2. PRELIMINARIES

This section provides general information about the research. Basic explanations are given in a simple way to make it easier for other researchers reading the thesis to comprehend the content. Within the chapter, recommendation systems, privacy issue, debiasing methods are discussed.

2.1. Recommender Systems

Numerous internet users have undoubtedly engaged in online shopping at some point. Take, for instance, the scenario where you log into an online platform to conduct your child's stationery shopping at the end of the summer vacation. You begin by adding various school-related items such as pencils, notebooks, and erasers to your shopping cart. After a while, the shopping platform starts recommending products you might need, such as glue, scissors, and a school bag. In this case, a company that sells hundreds of thousands of items in its store has begun to offer you personalized products. At this juncture, the system that determines which products are presented as a priority to a specific customer is referred to as a recommender system.

This scenario can be diversified across numerous domains in today's world. For instance, make decisions on which song to listen to from a vast pool of millions, which movie to watch, or which news article to read, all with the assistance of these software systems. What fundamentally distinguishes these systems from primitive ones based solely on popularity—recommending the most listened to songs, watched movies, or read news articles—is their ability to offer completely personalized recommendations. A prime example of this would be recommending a news article on economics that is the most read by all users to someone who only follows sports news.

The provision of recommendations that are not simply based on the most popular items but are instead personalized and tailored specifically to the user comes at a cost. This cost is met through data collected about the user, either implicitly or explicitly. This data is utilized in the creation of a user profile necessary for providing personalized information specific to the user. For example, direct feedback data types such as likes, ratings, and product reviews provided by users are considered *explicit* data. On the other hand,

data such as the user's previous purchases, pages they have browsed, and the identification of other users similar to them is classified as *implicit* data. Both types of data are employed in profile creation. The more enriched the profile becomes, the higher the quality and accuracy of the recommendations made. Fundamentally, this profile encompasses all activities that cause interaction between the user and the item. This is because the interests and tendencies of users are the most accurate indicators of future choices. Offering recommendations based on demographic data instead of *implicit* and *explicit* data types can be cited as an example of semi-personalized recommendations.

From the perspective of users, the diversification of services offered online and the extraordinary increase in consumption have necessitated not only the positioning of recommender systems but also the provision of high-quality recommendations. Users were finding it increasingly difficult to choose the most suitable options for themselves from the massive variety of items offered by these service providers or digital retailers. This selection process was time-consuming, exhausting, and often led to unsatisfactory outcomes. The overwhelming nature of this situation occasionally resulted in erroneous selections or decisions. Contrary to popular belief, having too many options began to decrease rather than increase user welfare. This revealed the truth that having more choices is not always better.

Recommender systems basically work around a user-item matrix. In the context of recommender systems, the term "*user*" refers to the consumer within the system, while "*item*" denotes the concept being offered or sold within the relevant domain. A recommender system operates based on the correlation between these two terms and is managed by the producer or service provider. An example user-item matrix can be seen below.

Table 2.1. Example of a User-Item Matrix in Recommendation Systems

	Item 1	Item 2	Item 3	Item 4
User 1	5	3		4
User 2	4			2
User 3		2	5	
User 4		4		5

In the realm of recommender systems, how does the mechanism proceed when it aims to recommend an item to a user, and what is its operational methodology? Essentially, the computations performed by these systems are fundamentally grounded in probability calculations, a branch under the broader umbrella of statistics. By employing various probabilistic methods, the system predicts the ratings that a user might assign to each item within the system. At the process's conclusion, items possessing the highest calculated rating values are sorted to the relevant user by the system. Over time, the process has evolved and diversified in response to changes in the domain of recommendation, customer expectations, and the needs of sellers, producers or service providers.

Recommender systems formulate their recommendations in two distinct phases. In the initial phase, they predict a representative rating for these recommendations. To elaborate further, for a user-item combination, the system predicts the rating value. This involves estimating the missing values in a user-item (denoted as u for users and i for items) matrix, essentially addressing a matrix completion problem. Algorithms related to this phase are developed specifically to solve this problem.

In the second phase of generating recommendations, recommender systems select the items to be recommended by using the matrix that has been filled with predicted representative values. The list of items to be recommended is conventionally referred to as k in the literature. The recommender system then presents this list of items to its user as a *top-k* list, effectively narrowing down the vast array of options to a curated selection most likely to be of interest to the user.

In the context of recommender systems, there are scenarios where, for instance, a user who listens to rock music may be recommended different songs from the rock genre, and at times, they might also receive recommendations for punk-metal music. This illustrates that while recommendations made through direct relationships are valuable, those generated by establishing indirect relationships can be even more beneficial. This approach underscores the system's capacity to discern and cater to the nuanced preferences of users, potentially enhancing user satisfaction by introducing them to a broader spectrum of music that aligns with their tastes, yet might not have been directly considered.

Recommender Systems represent a relatively newer suite of techniques that operate on a fuzzy logic basis, in contrast to the classical information retrieval methods employed by

search engines or databases. As consumption via digital platforms continues to rise, the interest in these systems is expected to increase proportionally. Nowadays, numerous institutions serving across various domains actively utilize these methodologies to enhance their services and user experience. This evolution highlights the growing significance of recommender systems in navigating the vast and expanding digital landscape, offering a more personalized and relevant selection of content to users.

Recommender systems fundamentally generate recommendations for users to consume, yet they perform a multitude of functions beyond this primary role. However, when considering the overarching objective, increasing product sales emerges as the foremost goal of recommender systems. These systems are developed with the intention of boosting sales volume and profitability. Therefore, these functions can be viewed as the facet of recommender systems that pertains to the producers/providers. They are managed and refined by them, underlining the systems' crucial role in driving commercial success and enhancing the strategic capabilities of businesses in various sectors. The functions of recommendation systems can be listed as follows;

- In the context of recommender systems, one of the paramount operational activities is recommending items that are *relevant* to the user. Users are much more likely to consume items that are not necessarily popular but are deemed interesting by them. While the level of relevance is a crucial activity, it alone is not sufficient.
- *Increasing the number of items* consumed appears to be a fundamental function of recommender systems at first glance. Examples of this include enhancing the basket size for an e-commerce portal or extending the duration of service usage for a streaming provider.
- Enhancing consumer *satisfaction* and *loyalty* is another critical aspect. When users consume an item recommended to them and find satisfaction in it, their trust in the seller/provider increases. Furthermore, a user whose trust has been gained is more likely to provide additional feedback to the system. This feedback allows for better profiling of the user and, consequently, better recommendations. This cycle encourages the user to engage more frequently with the system, thereby increasing their loyalty to it. This multifaceted approach underscores the strategic importance

of tailoring recommendations not just to drive sales, but also to foster a deeper, more engaged relationship between the consumer and the service or product provider.

- Understanding *consumer preferences* more accurately allows producers/providers to optimize manufacturing, inventory, and logistics costs by better analyzing user behaviors. This, in turn, facilitates more informed decisions regarding the conduct of marketing activities.
- *Serendipity* refers to the recommendation of unexpected items to users, which is distinct from *diversity* and *novelty*. The goal of this function is to surprise the user, a challenging yet highly rewarding endeavor when achieved, leading to extraordinary increases in user *satisfaction* and *loyalty*. The underlying principle is to introduce the user to a suitable product that transitions them from one *niche* product group to another, thereby offering a *discovery* experience. For instance, a user regularly consuming a specific item might be surprisingly recommended an unrelated product, such as recommending a sewing machine to someone who enjoys cheese, aiming not to introduce a new product but to uncover a new area of interest.
- Increasing the *variety of products* consumed aims to facilitate the selection of items by consumers that they have not yet searched for, are unaware of needing, or find challenging to discover without a pre-existing demand. In other words, a recommender system fulfilling this function should propose items not because they are popular or widely consumed/sold but because they have potential appeal. Practically, recommender systems recommend the best k items, expected to showcase a variety of different types rather than similar items. This diversity ensures users are less likely to dismiss lists containing items too alike, preventing boredom from repeated recommendations of similar items.

Currently, there are a number of technology companies that have been using recommender systems for a long time and actively, and beyond that, have implemented the above functions properly and have supported the development of this set of technologies. Recommendation systems can be analyzed through these organizations as follows.

GroupLens fundamentally acts as an idea recommendation system. After collecting

ratings from readers, it attempted to predict whether a user would like a new article when considering reading it. Therefore, it has been benefiting from recommender systems since their earliest stages. However, it later generated recommendations based on the *collaborative filtering* approach. The approach and techniques developed here were subsequently applied in book and movie domains, such as *BookLens* and *MovieLens*.

Amazon.com has established itself as one of the pioneering companies adopting recommender systems in the e-commerce sector. Initially focusing solely on book sales, the platform has gradually expanded to encompass a wide range of product categories, from electronics to software. *Amazon's* recommendation mechanisms are built upon users' *explicit* ratings, purchase, and browsing habits, offering them personalized recommendations. Ratings are collected on a scale ranging from one to five stars, and these recommendations are displayed on the homepage each time users log into their accounts. By developing models that integrate both *explicit* and *implicit* feedback, *Amazon* conducts a more comprehensive analysis of user behavior, enhancing the accuracy of its recommender systems. This approach combines information from both *explicit* and *implicit* data sets, providing users with more relevant and personalized recommendations.

Netflix initially started as a DVD rental service before transitioning to offer its content online, and currently provides movies and television programs on a subscription basis. Allowing users to rate content on a five-point scale, *Netflix* utilizes these ratings and viewing actions within its recommendation systems. It provides users with recommendations based on their viewing history, accompanied by explanations on why these recommendations are suitable for them. This strategy aids users in discovering new content and contributes to their decision-making processes. *Netflix's* recommendation system not only enhances user experience but also has the potential to increase customer loyalty and retention. Moreover, *Netflix* has made significant contributions to the research community by offering a platform for work on *collaborative filtering* algorithms through the *Netflix Prize* competition. This contest gave participants the opportunity to develop and test recommendation

algorithms using *Netflix* user movie ratings. *Netflix* provided both training and test datasets for the competition, enabling the evaluation of algorithms in real-world scenarios. The contest significantly contributed to the development of recommendation system algorithms, particularly in advancing hidden factor models, and drew attention to research in this area. The *Netflix Prize* is recognized as an initiative that underscored the importance of proper evaluation methods in the design of recommendation systems and led to innovations in the field.

Google News has developed a system that offers personalized news recommendations based on users' click history. During this process, user clicks are attributed to individual users through identification mechanisms activated via *Gmail* accounts, and these clicks are considered as *implicit* positive feedback towards news articles. Despite the absence of an opportunity for users to explicitly indicate their likes or dislikes towards articles, clicks are processed as indicators of user preferences. The collected *implicit* data is analyzed through *collaborative filtering* algorithms, resulting in the generation of personalized news recommendations that are likely to interest users. The details of this personalization approach, along with a description of a specific *collaborative filtering* system, are examined in more detail in the relevant literature.

Platforms like *Facebook*, a *social networking* site, offer potential friend recommendations to enhance users' social connections, a process motivated differently from product recommendations. While product recommendations aim to increase sales, directly benefiting the sellers' revenue, social connection recommendations aim to improve users' *social network* experiences, contributing to the platform's overall growth. Social networks rely on this growth to maximize advertising revenues. Potential friend recommendations address a domain known as the link prediction problem in social network analysis, by promoting the expansion of the network and encouraging more interactions among users. These recommendations are based on structural relationships rather than user preferences, hence the algorithms employed require a different approach compared to product recommendation systems.

Therefore, recommendation systems play an important role in the processes of organizations providing digital services. In some organizations, this role has a broad scope, while in others it is more limited. However, the functions of recommender systems will be enriched depending on the type and nature of the business.

Main Approaches in Recommender Systems

In the era of digital information overload, recommender systems have emerged as essential tools for filtering and personalizing content, products, and services to meet individual preferences and needs. These systems leverage various algorithms and methodologies to predict and recommend items that a user may find interesting or useful. Among the most prominent approaches are *collaborative filtering*, *content-based filtering*, *knowledge-based systems*, and *hybrid models*. *Collaborative filtering* analyzes user interactions and behaviors to identify patterns and recommend items liked by similar users. *Content-based filtering*, on the other hand, focuses on the properties of items themselves, recommending items similar to those a user has liked in the past. *Knowledge-based systems* utilize *explicit* knowledge about users and items to generate recommendations, often useful in scenarios with specific requirements or when historical interaction data is scarce. *Hybrid models* combine elements from multiple approaches to capitalize on their strengths and mitigate their weaknesses, aiming to provide more accurate and relevant recommendations. This paper delves into these fundamental approaches, discussing their principles, advantages, and limitations, and illustrating their application in various domains.

Content-Based Filtering (CBF) is a sophisticated approach within recommender systems, fundamentally relying on the analysis of item attributes and user preferences to generate personalized recommendations. It stands distinct from other recommendation strategies by focusing on the content of items—such as keywords, genres, or other specific attributes—thus aligning recommendations closely with the user’s past likes or interests. This method capitalizes on detailed descriptions of items and the construction of comprehensive user profiles, which can be enhanced through direct feedback or inferred from user interactions. Notably, CBF excels in offering highly personalized recommendations and adeptly addressing the new item challenge, as it does not require historical user-item

interactions but rather depends on the intrinsic characteristics of the items themselves. Despite its strengths, CBF faces challenges such as the potential for over-specialization, limiting recommendations to the user's established preferences and possibly hindering the discovery of diverse content. To counteract these limitations, advanced algorithms incorporating natural language processing and machine learning are increasingly employed, aiming to deepen the system's understanding of content and expand the breadth of recommendations. By leveraging these technologies, CBF systems strive to not only refine their predictive accuracy but also to enhance the variety and novelty of the recommendations they provide, thereby enriching the user experience in domains as varied as e-commerce, digital libraries, and online streaming platforms.

Knowledge-Based Filtering (KBF) represents a sophisticated approach in recommender systems, particularly effective in scenarios lacking extensive user interaction data, such as in the recommendation of newly introduced products or infrequently purchased items like consumer electronics. Unlike *collaborative* or *content-based* methods that rely on historical user-item interactions, KBF utilizes detailed domain knowledge about items and user preferences to craft recommendations. This approach is distinguished by its reliance on structured information regarding product specifications and quality, coupled with user-defined needs and criteria. In domains where customers typically make one-off purchases without a significant history of prior transactions, KBF systems engage in a dynamic dialogue with users, gathering preferences and prioritizing attributes through more advanced, conversational interactions. This method not only circumvents the cold start problem by not depending on past behaviors but also offers tailored recommendations by matching *explicit* user requirements with a comprehensive database of product attributes. While *knowledge-based* recommenders, including *constraint-based systems*, are adept at addressing personalization challenges in complex decision-making contexts, they require significant effort to construct and maintain an accurate and exhaustive knowledge base. The effectiveness of KBF hinges on the depth and precision of the underlying domain knowledge, marking it as a resource-intensive yet highly precise recommendation strategy, especially suitable for complex items and services.

Hybrid Approaches in recommender systems represent a strategic fusion of multiple recommendation techniques, aiming to harness the strengths and mitigate the limitati-

ons of individual approaches, thereby enhancing the *quality*, *diversity*, and *accuracy* of recommendations. This amalgamation, which can include any combination of *collaborative filtering*, *content-based filtering*, and *knowledge-based* strategies, alongside advanced machine learning models, offers a versatile solution to various challenges inherent in recommender systems. For instance, *hybrid systems* adeptly address the cold start problem by leveraging *content-based* or *knowledge-based* inputs when collaborative data is sparse, thus ensuring relevant recommendations for new users or items. Moreover, by integrating diverse data sources and recommendation logics, such as combining community insights with detailed item information, hybrid approaches can deliver more nuanced and personalized user experiences. This multifaceted strategy not only improves recommendation precision but also enhances user satisfaction across different domains by providing recommendations that are both accurate and reflective of a user's unique preferences and needs. The implementation of hybrid models varies, including but not limited to, the simultaneous use of different methodologies, the sequential application of techniques, or even the development of complex models that inherently blend several recommendation philosophies, showcasing the flexibility and creative potential of hybrid approaches in overcoming the constraints of singular recommendation systems.

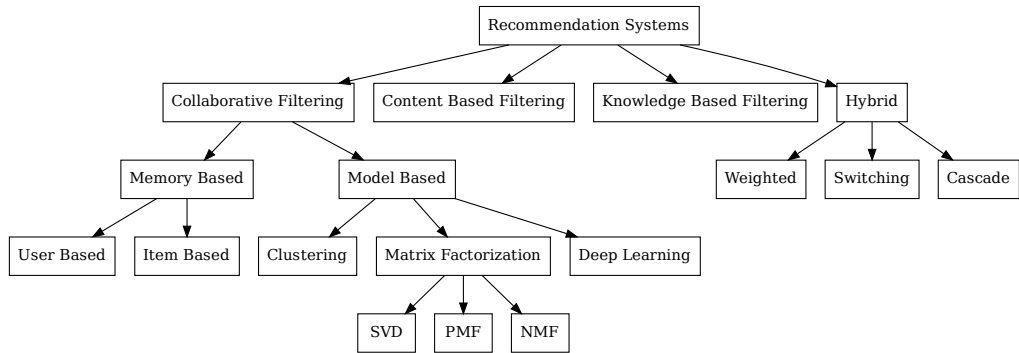


Figure 2.2. Main approaches in recommender systems

Collaborative Filtering (CF) is a pivotal mechanism in recommender systems, predicated on the principle that users with similar historical preferences are likely to have congruent interests in the future. This technique does not rely on specific information about the items but rather on the aggregation of user preferences and behaviors. CF bi-

furcates into *Memory-Based*, which calculates similarities among users or items to make recommendations, and *Model-Based* approaches, which use predictive algorithms, ranging from *matrix factorization* to deep learning, to deduce user preferences. Despite its simplicity and widespread adoption for its capacity to deliver personalized recommendations, CF confronts challenges such as scalability, data sparsity, and the cold start problem, where lack of data on new users or items hampers effective recommendation. Moreover, it may exhibit *popularity bias*, favoring frequently chosen items and potentially neglecting *niche* ones. Yet, its foundational premise—that shared historical preferences can guide future recommendations—remains a compelling rationale for its extensive use in crafting tailored user experiences.

In conclusion, the landscape of recommender systems is diverse, with each approach offering unique advantages and facing specific challenges. *Collaborative filtering*, *content-based filtering*, *knowledge-based systems*, and *hybrid approaches* represent the main strategies employed today. As technology advances, expect these methods to evolve further, offering more sophisticated and personalized recommendation capabilities.

2.2. Collaborative Filtering

People, being inherently social creatures, often base their consumption behaviors on the preferences of others. They place significant trust in recommendations made by others and are heavily influenced by them. For instance, they may read a book liked by their peers, watch movies recommended by their favorite film critic, or even purchase the same detergent their mother uses. This trust stems from a belief in the shared understanding within their social circles. The *collaborative filtering* approach is founded on the concept of artificially replicating this social circle with similar understandings, utilizing previously created user profiles to facilitate recommendations.

To put the concept of *collaborative filtering* into simpler terms, it revolves around examining profiles to identify users with similar past behaviors and simulating the process of these users making recommendations to each other. In other words, the collective behaviors of groups formed by users interested in similar items can enable the provision of recommendations to other users who have profiles similar to those of the group members.

This approach is grounded in a straightforward family of recommendation algorithms

known as neighborhood models, which belong to a broader class of models referred to as *collaborative filtering*. The term "*collaborative filtering*" denotes the collaborative use of ratings from multiple users to predict missing ratings. In practice, recommendation systems can be much more complex and data-rich, incorporating a wide variety of auxiliary data types. For instance, in content-based recommendation systems, content plays a primary role in the recommendation process, where users' ratings and the attribute descriptions of items are used to make predictions. The core idea is that user interests can be modeled based on the characteristics (or attributes) of items they have rated or accessed in the past. Another framework involves knowledge-based systems, where user interests are specified interactively, and user specification is combined with domain knowledge to provide recommendations. In more advanced models, contextual data such as temporal information, external knowledge, location data, social information, or network data can be utilized.

Collaborative filtering approaches leverage the collective input of user ratings to generate recommendations. A significant hurdle in the development of *collaborative filtering* techniques is the sparsity of the ratings matrix. To illustrate, let's consider a movie recommendation application where users rate movies to express their preferences. Given the vast array of movies available, it's common for users to have rated only a small subset. Consequently, the majority of potential ratings remain unfilled. These filled-in ratings are known as observed ratings. In this text, "*specified*" and "*observed*" ratings are terms that will be used synonymously. Ratings that have not been filled in will be described as "*unobserved*" or "*missing*."

The foundational concept behind *collaborative filtering* methods hinges on the observation that ratings provided across various users and items often exhibit significant correlation, enabling the prediction of unassigned ratings. Consider, for example, two users, named *A* and *B*, whose preferences are remarkably similar. If their provided ratings closely align, such similarities can be detected by the underlying algorithm. In scenarios where only one has rated an item, it's likely that their ratings would be similar. This similarity can be employed to infer about values that are not explicitly provided. Most models dedicated to *collaborative filtering* focus on leveraging either item-item correlations or user-user correlations for prediction purposes. Some models utilize both types of corre-

lations. Moreover, certain models employ carefully designed optimization techniques to create a training model, similar to how a classifier constructs a training model from labeled data. This model is then used to assign values to missing entries in the matrix, akin to how a classifier predicts missing test labels. There are primarily two types of approaches used in *collaborative filtering*: *memory-based* methods and *model-based* methods.

1. ***Memory-based methods***: This method also known as *neighborhood-based collaborative filtering* algorithms, represent some of the initial approaches to *collaborative filtering*. In these methods, predictions for user-item pair ratings are made based on the concept of neighborhoods. The definition of neighborhoods can be approached from two distinct perspectives:

- *User-based collaborative filtering*: In this scenario, recommendations for a specific user, referred to as target user A , are derived from the ratings given by users who share similar preferences or tastes. The core principle involves identifying users who resemble target user A in terms of their ratings and leveraging the weighted averages of these similar users' ratings to infer the ratings for items not yet rated by A . For instance, if A and B have historically shown a pattern of rating movies similarly, A 's given ratings for the movie "*Terminator*" could be instrumental in forecasting B 's yet-to-be-given ratings for the same movie. Typically, the ratings from the top k users who are most closely aligned with B 's preferences are utilized for predicting B 's ratings. The process involves calculating similarity scores between users, which are essentially computed between the rows of the ratings matrix, to identify users with similar tastes.
- *Item-based collaborative filtering*: To forecast how user A will rate a specific item, B , the initial action is to identify a collection, S , of items that bear the closest resemblance to the target item B . The ratings for the items within set S , as provided by user A , serve as the basis for predicting user A 's potential liking for item B . Therefore, B 's ratings in similar science fiction movies like *Last Samurai* can be used to predict his rating in *Troy*. This process entails the calculation of similarity measures between the columns of the ratings matrix,

which aids in identifying items that are similar..

Memory-based methods boast the advantage of being straightforward to implement, coupled with the fact that the recommendations they produce can often be easily justified. However, these algorithms tend to struggle in the context of sparse ratings matrices, meaning they may falter when it's challenging to identify users sufficiently akin to a given user, such as B , who have provided ratings for a specific item like *Last Samurai*. This difficulty can lead to a lack of robustness in predicting how B would rate *Last Samurai*, indicating that these approaches might not always offer comprehensive coverage in terms of rating predictions. Despite this, the issue of coverage often becomes negligible when the focus is on identifying only the *top-k* recommended items.

2. ***Model-based methods***: Within the framework of *model-based approaches*, techniques from *machine learning* and *data mining* are employed to construct predictive models. When these models are characterized by parameters, the determination of these parameters occurs through an optimization framework. This category encompasses various methodologies, including *decision trees*, *rule-based models*, *Bayesian methods*, and *latent factor models*. Notably, certain methods among these, for instance, *latent factor models*, exhibit a significant capability for coverage, making them effective even in scenarios where the ratings matrices are sparse.

While *memory-based collaborative filtering* algorithms are appreciated for their straightforwardness, their heuristic approach often limits their effectiveness across various scenarios. The differentiation between *memory-based* and *model-based* methods, however, can be seen as somewhat contrived since *memory-based* techniques can also be viewed as models based on similarity, though they are heuristic in nature. Recently, it has been demonstrated that integrating *memory-based* with *model-based* methods can yield highly accurate outcomes.

Utilized Recommendation Algorithms

Recommendation systems are pivotal in filtering vast amounts of information to recommend items that are most likely of interest to the user. Among the various approaches,

prediction algorithms play a crucial role in determining the relevance of items to a user's preferences. This thesis outlines some of the most commonly used prediction algorithms in recommendation systems.

- ***Most Popular (MP)***: *MP* algorithm is a simple yet highly effective approach used in recommendation systems, which ranks items based on their popularity, determined by the number of interactions or ratings they receive from users. Unlike personalized recommendation algorithms that tailor recommendations to individual user preferences, the *MP* algorithm assumes that items preferred by the majority will likely be appreciated by a new or average user as well. This method is particularly useful as a baseline model or in scenarios where user-specific data is scarce or unavailable. Despite its simplicity, the *Most Popular* algorithm can often provide surprisingly relevant recommendations, especially in contexts where trending or widely acclaimed items are of interest to a broad audience.
- ***Item Average (IA)***: *IA* algorithm is a straightforward yet effective approach for recommendation systems, focusing on calculating the average rating or interaction level for each item across all users who have interacted with it. This method assumes that items with higher average ratings are generally preferred by users, making it a simple baseline for recommendations. Despite its simplicity, *IA* can provide a useful starting point for recommendation systems, especially in cases where more complex models may not significantly outperform simpler approaches due to sparse data. The *Item Average* algorithm highlights the importance of understanding item popularity and general user consensus, serving as a foundational metric from which more sophisticated recommendation algorithms can build upon.
- ***Maximum Margin Matrix Factorization (MMMF)***: *MMMF* is a sophisticated algorithm for recommendation systems that integrates the principles of *matrix factorization* with margin optimization, similar to *support vector machines*. *MMMF* seeks to discover latent factors for users and items by factorizing the user-item interaction matrix, while also maximizing the margin between positive and negative interactions. This approach allows *MMMF* to not only predict the presence of an interaction but also to confidently differentiate between users' likes and dislikes. By focusing

Table 2.2. Detailed information about utilized recommendation algorithms

No	Family	Recommendation Algorithm	Abbreviation
1	Non-personalized	Most Popular	MP
2	Non-personalized	Item Average	IA
3	Matrix factorization-based	Maximum Margin Matrix Factorization (Weimer et al. 2008)	MMMF
4	Matrix factorization-based	Weighted Matrix Factorization (Hu et al. 2008)	WMF
5	Matrix factorization-based	Hierarchical Poisson Factorization (Gopalan et al. 2015)	HPF
6	Probabilistic	Weighted Bayesian Personalized Ranking (Gantner et al. 2012)	WBPR
7	Probabilistic	Indexable Bayesian Personalized Ranking (Le and Lauw 2017)	IBPR
8	Clustering-based	Spherical k -means (Salah et al. 2016)	SKM
9	Neural networks-based	Neural Matrix Factorization (He et al. 2017)	NEUMF
10	Neural networks-based	Variational Autoencoder for CF (Liang et al. 2018)	VAECF

on the margin, *MMMF* effectively enhances the discrimination power of the model, leading to more accurate and robust recommendations. The incorporation of margin optimization makes *MMMF* particularly effective in scenarios where it's crucial to distinguish strongly between preferred and non-preferred items, providing a clear advantage in precision-focused recommendation tasks.

- **Weighted Matrix Factorization (WMF):** *WMF* is a *collaborative filtering* algorithm that enhances traditional *matrix factorization* techniques by introducing weights to the user-item interaction matrix. This approach assigns different weights to the interactions, reflecting the confidence in observed data versus missing data. In the context of *implicit* feedback datasets—where interactions are indicators of preference but the absence of interaction is not necessarily an indicator of disinterest—*WMF* effectively distinguishes between the significance of observed and unobserved interactions. By doing so, *WMF* can better model the underlying preferences of users, leading to more accurate and personalized recommendations. The weighting scheme allows the algorithm to focus more on reliable interactions, improving the recommendation quality in sparse and noisy datasets.
- **Hierarchical Poisson Factorization (HPF):** *HPF* is an advanced algorithm for recommendation systems that models user-item interactions using a *Poisson distribution*, which is particularly suited for count data, such as the number of times a user has interacted with an item. *HPF* introduces a hierarchical structure with gamma

priors over the latent factors, enabling the model to capture both the sparsity and the over-dispersion commonly observed in real-world interaction data. This hierarchical approach allows *HPF* to effectively infer the underlying preferences and behaviors of users, even with very sparse data. The flexibility and robustness of *HPF* make it an excellent choice for recommendation scenarios where understanding the intensity of interaction is as important as the interaction itself, such as in content recommendation and e-commerce environments.

- ***Weighted Bayesian Personalized Ranking (WBPR)***: *WBPR* is an enhancement of the *Bayesian Personalized Ranking (BPR)* framework, designed to improve recommendation systems by incorporating weights into the pairwise ranking optimization process. *WBPR* assigns different weights to user-item interactions based on their importance or confidence levels, allowing the model to focus more on significant interactions while still considering less confident ones. This approach enables a more nuanced differentiation between preferred and non-preferred items, leading to more accurate and personalized recommendations. The weighted scheme in *WBPR* is particularly effective in dealing with *implicit* feedback datasets, where the varying degrees of interaction strength can be leveraged to better understand user preferences.
- ***Indexable Bayesian Personalized Ranking (IBPR)***: *IBPR* is a variant of the Bayesian Personalized Ranking algorithm that is designed for scalability and efficiency in large-scale recommendation systems. Unlike traditional *BPR*, which focuses on pairwise ranking of items based on user preferences, *IBPR* incorporates an indexing mechanism. This mechanism allows for faster retrieval of item rankings by reducing the computational complexity involved in comparing every possible item pair. By efficiently organizing the data, *IBPR* enables real-time recommendation scenarios where quick and dynamic ranking of items is crucial. This approach not only improves the user experience by providing timely recommendations but also enhances the system's ability to handle vast datasets without compromising on recommendation quality.
- ***Spherical k-Means (SKM)***: *SKM* is a variation of the traditional *k-Means clustering*

algorithm, designed specifically for clustering data on the surface of a sphere. This algorithm is particularly effective for high-dimensional and sparse data, such as text documents represented in a term frequency-inverse document frequency (*TF-IDF*) format, where *cosine similarity* measures are more meaningful than *Euclidean* distances. In *Spherical k-Means*, the similarity between data points is measured based on the angle between them, treating each point as a vector from the origin. This approach ensures that the clustering is based on the direction of the data points rather than their magnitude, making it highly suitable for applications in information retrieval and text mining, where the focus is on the pattern of term occurrence rather than the absolute term counts.

- **Neural Matrix Factorization (*NeuMF*)**: *NeuMF* is an advanced recommendation system algorithm that combines traditional *matrix factorization* techniques with neural network architectures to enhance the prediction of user-item interactions. By integrating the linearity of *matrix factorization* for capturing general user-item interactions and the non-linearity of neural networks for modeling complex and non-obvious user-item relationships, *NeuMF* effectively captures a wide range of interaction patterns. This hybrid approach allows *NeuMF* to outperform either approach used in isolation, particularly in scenarios where the data exhibits both simple and complex interaction patterns. The flexibility and power of *NeuMF* make it a popular choice for modern recommendation systems seeking to provide highly accurate and personalized recommendations.
- **Variational Autoencoder for Collaborative Filtering (*VAECF*)**: *VAECF* is a state-of-the-art recommendation algorithm that utilizes the principles of variational autoencoders (VAEs) to learn latent representations of users or items for *collaborative filtering*. By encoding user-item interaction data into a lower-dimensional latent space and then reconstructing the input from this space, *VAECF* is able to capture the complex, non-linear relationships inherent in the data. This approach allows for more nuanced and accurate recommendations compared to traditional *matrix factorization* methods. Additionally, *VAECF* handles missing data elegantly, making it particularly well-suited for sparse datasets common in recommendation systems.

The probabilistic nature of VAEs also provides a measure of uncertainty in the recommendations, offering insights into the model's confidence in its predictions.

2.3. Privacy Concerns in Recommender Systems

Privacy concerns in recommender systems have become increasingly prominent as these systems gain ubiquity in various domains, including e-commerce, streaming services, and social media platforms. At their core, recommender systems aim to provide personalized recommendations to users by analyzing their past behavior, preferences, and interactions. While the personalization benefits are clear, these systems raise significant privacy concerns because they inherently require access to sensitive and personal user data. This discussion explores the nature of privacy concerns within recommender systems, focusing on the types of data involved, potential privacy risks, and the broader implications for users and society.

Recommender systems rely on a wide array of user data, including *explicit* preferences (e.g., *ratings, likes*), *implicit* signals (e.g., *browsing history, purchase history*), and personal information (e.g., *age, location*). This data is critical for the system to function effectively; however, it also represents a potential goldmine for malicious actors if not properly secured. The primary privacy concern is the unauthorized access and misuse of this personal data. Cybersecurity breaches can expose user data, leading to identity theft, financial fraud, and an erosion of user trust.

Moreover, even when data is not maliciously accessed, the very operation of recommender systems can lead to privacy invasions. For example, through inference algorithms, systems can deduce and reveal sensitive information about a user's *political views, health status, or personal preferences* that the individual might prefer to keep private. This aspect of "data inference" poses a subtle but significant privacy risk, as users may not be aware of how much the system can learn about them beyond what they explicitly shared.

Another significant privacy concern is the extent of behavioral tracking and profiling conducted by recommender systems. To refine their recommendations, these systems often track user behavior across multiple platforms and devices, creating detailed profiles of individual users. While beneficial for personalization, this practice raises questions about user consent and control over their data. Users may not fully understand what data is be-

ing collected, how it is being used, or with whom it is being shared, leading to concerns over a lack of transparency and autonomy.

From a broader societal perspective, recommender systems can inadvertently contribute to the creation of *echo chambers* and *filter bubbles*, where users are exposed only to content that aligns with their existing views or preferences. While not a direct privacy concern, this phenomenon relates to privacy insofar as it is a consequence of personalized algorithms leveraging private data to curate content. The resulting lack of exposure to diverse viewpoints can reinforce biases and contribute to societal polarization, underscoring the complex interplay between personalization, privacy, and the public good.

The privacy concerns associated with recommender systems have prompted regulatory responses, such as the *General Data Protection Regulation (GDPR)* in the European Union, which aims to give individuals more control over their personal data. These regulations introduce requirements for consent, transparency, and data minimization, among others, directly addressing some of the privacy issues raised by recommender systems. However, compliance and enforcement remain challenging, as the technology and methodologies behind these systems continue to evolve rapidly.

Ethically, the development and deployment of recommender systems raise questions about the balance between *personalization benefits* and *privacy risks*. Developers and operators of these systems face the task of designing algorithms that respect user privacy while still delivering the personalized experiences that users value. This includes considerations around data collection practices, informed consent, and the implementation of mechanisms that allow users to control their data and understand how it is used.

In conclusion, while recommender systems offer significant benefits in terms of personalization and user experience, they also introduce a range of privacy concerns that must be carefully managed. These concerns encompass the collection and use of sensitive data, the risks of unauthorized access and data inference, the implications of behavioral tracking and profiling, and the broader societal impacts related to *echo chambers* and *filter bubbles*. Addressing these concerns requires a concerted effort from system designers, operators, regulators, and users themselves to ensure that the advancement of recommender technologies does not come at the expense of individual privacy and societal well-being.

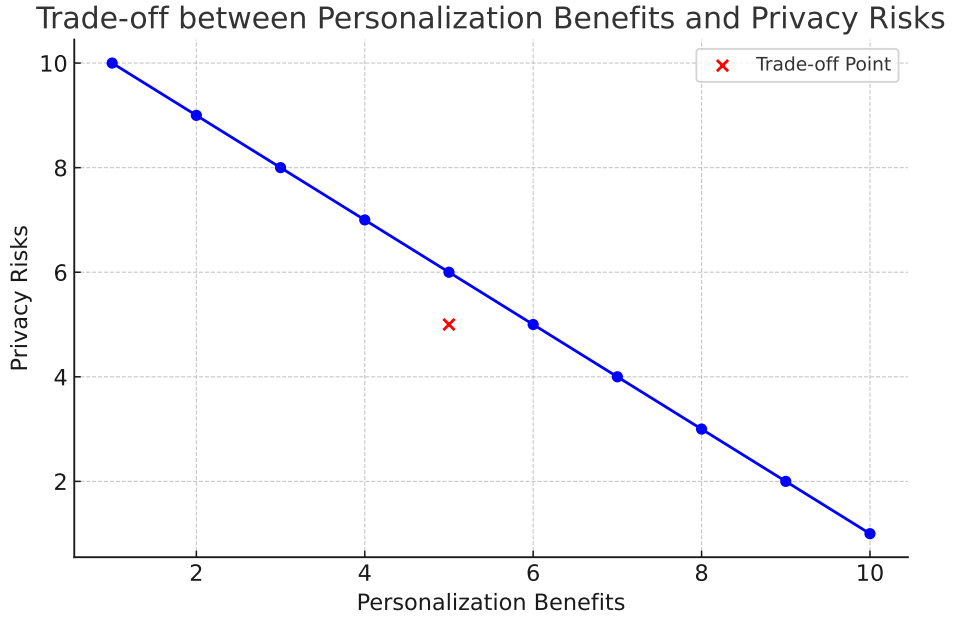


Figure 2.3. Personalization Benefits and Privacy Risks Trade-off

2.3.1. Methodologies for Enhancing Privacy

Enhancing privacy in recommender systems is a critical concern that has attracted significant attention from the researchers. Various methodologies have been developed to address the privacy challenges inherent in designing and deploying recommender systems. These methodologies focus on minimizing the risk of personal data exposure without significantly compromising the quality of the recommendations provided. This discussion delves into several key approaches, including *data anonymization*, *differential privacy*, *federated learning*, and *encryption techniques*, as underpinned by academic research.

Data anonymization, in recommendation systems involves the process of *obfuscating* or removing personally identifiable information from the datasets used to train these systems. The goal is to ensure that the data cannot be traced back to individual users, thereby protecting their privacy. This technique can involve methods such as pseudonymization, where users are assigned random identifiers, or aggregation, where individual data points are combined into larger datasets to obscure personal details. While *data anonymization* can significantly reduce privacy risks, it must be carefully implemented to avoid diminis-

hing the quality of the recommendations due to loss of useful information. In the context of ensuring user privacy within recommendation systems, *data anonymization* serves as a pivotal technique, particularly through the implementation of *perturbation* and *obfuscation* methods. *Perturbation* involves the deliberate addition of noise to the data or the modification of actual data values to prevent the accurate identification of individual users, thus preserving their anonymity while maintaining the overall data structure for effective recommendation generation. *Obfuscation*, on the other hand, seeks to obscure sensitive information by generalizing or altering data attributes in a way that reduces the precision of the data, thereby mitigating the risk of re-identification. Both approaches aim to strike a balance between safeguarding user privacy and retaining sufficient data utility, highlighting the nuanced trade-offs that recommendation systems must navigate to deliver personalized experiences in a privacy-conscious manner.

Differential privacy provides a framework for quantifying the privacy guarantees of algorithms used in recommendation systems. It ensures that the output of an algorithm does not substantially change when any single individual's data is added or removed from the dataset, thereby protecting the privacy of users. In the context of recommendation systems, differential privacy can be applied by adding noise to the data or to the algorithm's outputs, ensuring that individual user preferences cannot be inferred. This approach allows for the generation of personalized recommendations while maintaining a high level of user privacy.

Federated learning is a decentralized approach to machine learning where the model is trained across multiple devices or servers without exchanging raw data. In recommendation systems, this means that personal data stays on the user's device, and only model updates are shared with the server. This technique significantly enhances privacy by reducing the amount of personal data that needs to be transmitted and stored centrally. Moreover, *federated learning* can enable personalized recommendations by leveraging data from many users while still respecting individual privacy.

Encryption techniques in recommendation systems are used to protect the privacy of user data during transmission and storage. *Homomorphic encryption*, for instance, allows for operations to be performed on *encrypted data* without needing to *decrypt* it, enabling the generation of recommendations based on *encrypted* user preferences. Secure multi-

party computation is another technique that allows different parties to jointly compute a function over their inputs while keeping those inputs private. These encryption methods ensure that sensitive user data remains confidential, providing a secure foundation for privacy-preserving recommendation systems.

In conclusion, the methodologies for enhancing privacy in recommender systems are diverse and under continuous development, reflecting the complex balance between personalization and privacy. These approaches, grounded in rigorous academic research, offer promising paths forward in addressing privacy concerns. However, they also underscore the need for ongoing innovation and adaptation as recommender systems evolve and the landscape of digital privacy continues to change.

2.3.2. Randomized Perturbation and Obfuscation

Randomized perturbation stands as a cornerstone technique in the domain of *privacy-preserving* recommender systems, aimed at safeguarding user data against unauthorized access and inference attacks. By integrating randomness into the dataset, this method delicately obscures individual user preferences and activities, rendering the extraction of personal information considerably more challenging for malicious entities.

The essence of *randomized perturbation* involves the deliberate alteration of rating values through the addition or subtraction of random noise. This process is executed in such a manner that the original distribution and statistical properties of the dataset are preserved to the greatest extent possible, ensuring the recommender system's ability to generate meaningful recommendations remains unaffected. The noise introduced is typically derived from a predefined probability distribution, with the *Gaussian (Normal) distribution* being a common choice due to its natural fit with the bell curve of many real-world datasets.

A pivotal challenge in implementing *randomized perturbation* is striking an optimal balance between privacy enhancement and the preservation of data utility. Excessive noise can significantly degrade the quality of the recommendations provided, leading to a diminished user experience. Conversely, insufficient perturbation may fail to adequately protect user privacy, leaving sensitive information vulnerable to inference attacks. Advanced techniques, such as differential privacy, have been proposed to quantify and control

this trade-off more effectively, providing a mathematical guarantee of privacy.

One of the primary advantages of *randomized perturbation* is its straightforward implementation and compatibility with existing recommender system architectures. It does not require substantial modifications to the recommendation algorithms themselves but rather applies a *pre-processing* step to the data. However, the technique’s efficacy is inherently tied to the chosen noise parameters, necessitating careful calibration to align with specific privacy requirements and dataset characteristics.

Obfuscation in the context of *privacy-preserving recommender* systems entails the deliberate complication of data to prevent the accurate attribution of actions or preferences to specific users. This technique introduces ambiguity into the dataset, effectively masking the true intentions or interests of individuals, thereby complicating efforts by unauthorized parties to construct detailed user profiles.

Obfuscation can be implemented through various means, such as the insertion of fictitious ratings, the modification of existing ratings, or the strategic blending of user profiles. Each approach serves to dilute the clarity of user-specific data, embedding a layer of uncertainty that protects individual privacy. The selection of obfuscation strategies is often dictated by the specific architecture of the recommender system and the nature of the underlying data.

The effectiveness of *obfuscation* is contingent upon its capacity to introduce sufficient ambiguity without fundamentally undermining the integrity of the recommendation process. This necessitates a nuanced approach to the design and deployment of *obfuscation* mechanisms, ensuring they are sufficiently *robust* to thwart privacy breaches while maintaining the overall relevance and personalization of the recommendations provided. Techniques such as clustering users based on broad preferences before applying *obfuscation* within these clusters can help maintain the balance between privacy and recommendation accuracy.

A notable challenge associated with obfuscation lies in its potential to inadvertently lead to a homogenization of the recommendations, particularly if the obfuscation is not carefully calibrated. This can result in a loss of personalization, diminishing the value proposition of the recommender system. Furthermore, the dynamic nature of user interactions and preferences necessitates ongoing adjustments to obfuscation strategies, ensuring

they remain effective against evolving privacy threats and continue to reflect the changing landscape of user data.

Both *randomized perturbation* and *obfuscation* offer valuable strategies for enhancing privacy in recommender systems, each addressing distinct aspects of the privacy preservation challenge. While *randomized perturbation* focuses on the statistical alteration of data, *obfuscation* targets the structural ambiguity of user information. Together, these methodologies form a comprehensive privacy-preserving framework, enabling recommender systems to deliver personalized experiences while steadfastly protecting user privacy.

2.3.3. The Long-Tail Model

The concept of the "*long-tail*" in the context of recommender systems is a fascinating aspect of modern digital commerce and content distribution, fundamentally altering how products and services are consumed and recommended. Originating from Chris Anderson's seminal work in 2004, the *long-tail* concept describes the shift away from a focus on a small number of mainstream, *popular items (the "head")* towards a vast number of *niche, less popular items (the "tail")* in the distribution curve of product sales or content popularity.

In traditional retail and media industries, the focus has been predominantly on the *head*, driven by physical shelf space limitations and the high costs of distribution, which necessitated concentrating on the most universally appealing and hence profitable items. However, the advent of the internet and digital technologies has dramatically reduced these constraints, enabling businesses to offer a much wider variety of items, including those that appeal to only a small segment of the market. This is where the *long-tail* finds its significance in recommender systems.

Recommender systems, powered by sophisticated algorithms, play a pivotal role in making the *long-tail* accessible and beneficial to both consumers and providers. They analyze vast amounts of data on user preferences and behavior to recommend items that users are likely to enjoy but might not have discovered on their own. This capability is especially crucial for items in the *long-tail*, which, despite their lower individual popularity, can cumulatively represent a significant portion of market demand.

For consumers, the *long-tail* enhanced by recommender systems means a more perso-

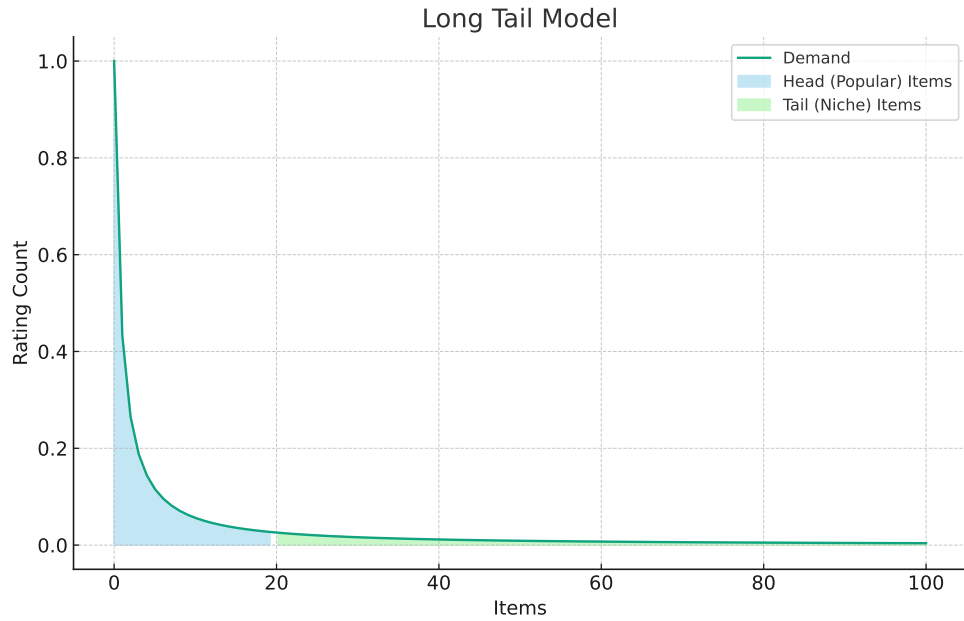


Figure 2.4. The Long-Tail Model

nalized shopping or content consumption experience. They are no longer limited to what is popular or available in local stores. Instead, they can explore a much broader range of products, including *niche* items that closely match their interests and preferences. This democratization of access significantly enriches the consumer experience, fostering a more diverse and satisfying consumption landscape.

From the perspective of content providers and retailers, the *long-tail* offers opportunities to cater to more specific customer segments and differentiate their offerings from competitors. By leveraging recommender systems to effectively surface *long-tail* items to the right audience, businesses can tap into previously untapped markets, increasing their customer base and enhancing loyalty through personalized engagement. Moreover, the *long-tail* can lead to improved inventory management and efficiency, as digital distribution reduces the need to predict which few items will be blockbusters, instead offering a wider selection with less risk of overstocking.

The economic implications of the *long-tail* are profound. It challenges the traditional 80/20 rule (*Pareto Principle*), which recommends that 80% of sales come from 20% of products. In the digital realm, the *long tail* theory posits that businesses can achieve significant revenues from the cumulatively large number of sales of *niche* items, potentially

rivaling or exceeding sales from the head. This shift requires businesses to rethink strategies around product range, marketing, and customer engagement, emphasizing the importance of data analytics and recommendation algorithms in uncovering and capitalizing on *long-tail* opportunities.

However, the exploitation of the *long-tail* through recommender systems is not without challenges. Ensuring the accuracy and effectiveness of recommendations requires sophisticated algorithms capable of dealing with sparse data and the inherent diversity of *long-tail* items. Additionally, there's the need to maintain a balance between recommending popular items and surfacing *niche* content, to cater to a wide range of user preferences without overwhelming them with too much choice.

The *long-tail* concept in recommender systems signifies a shift towards more personalized, diverse, and inclusive content and product recommendations. It reflects the changing dynamics of consumer demand and the power of digital technologies to satisfy a broader spectrum of preferences. As recommender systems continue to evolve, their ability to harness the potential of the *long-tail* will play a crucial role in shaping the future of digital commerce and content distribution, offering unique opportunities and challenges for businesses and consumers alike.

2.4. Popularity Bias Problem and Debiasing Methods

In the context of recommender systems, the concept of *popularity bias* necessitates an in-depth understanding of the dynamics between *popular* and *niche* items, which is fundamentally shaped by the *long-tail model*. This model has profound implications for how items are recommended to users and, consequently, for the diversity and personalization of user experiences in digital platforms.

Recommender systems, integral to content platforms like streaming services, e-commerce websites, and social media, aim to personalize the user experience by recommending items (products, movies, songs, etc.) based on individual preferences and behavior. However, these systems often gravitate towards recommending items that are already popular, creating a cycle where popular items gain more visibility, thereby becoming even more popular. This phenomenon is referred to as *popularity bias*.

Popularity bias in recommender systems can lead to a narrowing of user experiences,

where a small set of *popular items* dominates the recommendations at the expense of a vast number of *niche items* that make up the *long-tail*. The *long-tail* concept highlights that while a small number of items may account for the bulk of consumption (the head), there is a long distribution of items that, although each may account for a small part of the market share individually, collectively represent a significant portion of the market.

The challenge with popularity bias is multifold:

1. ***Diminished Discoverability:*** *Niche items* with potentially high personal relevance to individual users struggle to surface against the overwhelming visibility of *popular items*.
2. ***Reduced Personalization:*** The essence of personalization is compromised when recommendations are skewed towards widely popular items, neglecting the unique preferences and interests of users.
3. ***Impaired Long-Tail Utilization:*** The economic and cultural value hidden in the *long-tail* is underutilized, as the items therein fail to reach their potential audience.

Understanding the multifaceted challenge of *popularity bias* - spanning from diminished discoverability of *niche items*, through reduced personalization, to impaired *long tail* utilization - underscores the critical necessity for implementing *popularity debiasing* strategies in recommendation systems. Addressing these issues not only promises to enhance the visibility of *niche items* but also ensures that personalization truly reflects the diverse tastes and preferences of the user base. By actively working towards mitigating the dominance of *popular items*, systems can unlock the full potential of *the long-tail*, thereby democratizing content discovery and enriching the user experience. This endeavor to recalibrate the balance between *popular* and *niche items* not only aligns with the ethos of equitable visibility but also stimulates a more vibrant, diverse, and culturally rich digital ecosystem. In conclusion, the strategic implementation of *popularity debiasing* is pivotal in transitioning from a homogenized recommendation landscape to one that celebrates and elevates the unique and varied interests of each user, fostering a truly personalized and inclusive digital space.

Popularity bias in recommender systems is a pervasive issue that skews the distribution of recommended items towards those that are already popular, often at the expense

of *long-tail*, *niche* items that might be equally or more relevant to certain users. This phenomenon not only affects the diversity of content that is surfaced to users but also has broader implications for content creators and the overall digital ecosystem. In addressing this challenge, it is crucial to delve into the mechanisms of recommender systems, the implications of *popularity bias*, and potential strategies for mitigation.

Recommender systems leverage user data, including past behavior and preferences, to predict and recommend items that a user might find interesting or useful. These systems employ a variety of algorithms, including *collaborative filtering*, *content-based filtering*, and *hybrid methods*, to generate recommendations. Despite their effectiveness in personalizing user experiences, these algorithms can inadvertently amplify the visibility of *popular items* due to their reliance on user interaction data. *Popular items*, by virtue of being more frequently rated or interacted with, are more likely to be recommended, creating a feedback loop that further entrenches their dominance.

The consequences of *popularity bias* extend beyond just a lack of variety in recommendations. It can lead to a homogenization of content consumption, where *niche* and potentially more relevant items are marginalized. This not only diminishes the user experience by limiting exposure to a wider range of content but also impacts the economic viability of producing diverse content. For content creators, especially those producing *niche* or specialized content, this bias can significantly reduce visibility and accessibility to their intended audience, thereby affecting their motivation and ability to create.

Moreover, the dominance of *popular items* in recommendations can stifle *innovation* and *creativity* in content production. When success is increasingly determined by visibility within recommender systems, content creators might be incentivized to conform to prevailing trends rather than explore unique or innovative ideas. This can lead to a cultural impoverishment where less mainstream voices and perspectives are drowned out in a sea of popularity-driven content.

Addressing *popularity bias* requires a multifaceted approach that involves refining recommender algorithms to better capture and reflect the *diversity* of user interests and the breadth of available content. Techniques such as diversification strategies, which explicitly incorporate variety into the recommendation process, and re-ranking methods, which adjust recommendation lists to ensure a representation of *niche* items, have shown pro-

mise. Additionally, incorporating user feedback mechanisms that allow users to indicate their interest in discovering less popular content can help tailor recommendations to include a broader spectrum of items.

Another promising avenue is the development of algorithms that are inherently less susceptible to *popularity bias*. For example, algorithms that focus on item *novelty* and user curiosity can help surface less known items by prioritizing aspects other than historical interaction frequency. Likewise, leveraging machine learning techniques to detect and correct for bias in recommendation algorithms can help ensure a more balanced distribution of recommended items.

In conclusion, while *popularity bias* in recommender systems presents a significant challenge, it also offers an opportunity to rethink how these systems are designed and operated. By embracing strategies that promote diversity and balance, recommender systems can better serve the full spectrum of user interests and support a richer, more vibrant digital content landscape. This not only enhances the user experience but also contributes to a more equitable and dynamic cultural ecosystem, where all content, regardless of its inherent popularity, has the opportunity to reach its audience.

Addressing popularity bias requires deliberate algorithmic strategies within recommender systems:

- ***Diversification Techniques***: Algorithms can be designed to intentionally include a mix of *popular* and *long-tail items* in recommendations, ensuring that users are exposed to a broader spectrum of content.
- ***Personalization Enhancements***: Improving the sophistication of personalization algorithms to better understand and predict individual user preferences can help surface more *niche items* relevant to each user.
- ***Explicit Long-Tail Promotion***: Recommender systems can be programmed to specifically promote *long tail* items, either by boosting their recommendation likelihood or by featuring dedicated sections for *niche content*.
- ***Feedback Loop Adjustments***: Modifying the system to counteract the self-reinforcing feedback loops that *amplify popularity bias*, by ensuring that the system periodi-

cally reassesses and adjusts the weight given to user interactions with both *popular* and *niche* items.

In essence, addressing *popularity bias* in recommender systems is not just about enhancing the accuracy of recommendations but is crucial for fostering a diverse, vibrant digital ecosystem where all items, regardless of their inherent popularity, have the opportunity to reach their relevant audience. This endeavor supports the democratization of content consumption, allowing for a richer, more personalized user experience that values the depth and breadth of available content.

Popularity debiasing in recommender systems is an essential area of research aimed at mitigating the influence of *popularity bias* on the recommendations provided to users. *Popularity bias* occurs when items that are already popular are more likely to be recommended by the system, leading to a reinforcement loop that marginalizes *niche* or less popular items. This bias can diminish the diversity of the recommended items and affect user satisfaction by not accurately reflecting the wide array of user interests. To address this challenge, researchers have developed various *debiasing* techniques, categorized into *pre-processing*, *in-processing*, and *post-processing* methods, each with its unique approach to mitigating *popularity bias*.

Pre-processing methods focus on adjusting the input data before it is fed into the recommendation algorithm. The goal is to reduce the inherent biases in the data, thus preventing the recommender system from amplifying these biases in its output. One common approach is to re-weight the ratings in the dataset to increase the influence of ratings for less popular items and decrease the influence of ratings for popular items. Another technique involves augmenting the dataset with synthetic data for underrepresented items to balance the distribution of item popularity. These methods aim to create a more balanced dataset that reflects a wider range of items, thereby allowing the recommendation algorithm to discover and recommend *niche* items more effectively.

In-processing methods integrate debiasing directly into the recommendation algorithm. These approaches modify the algorithm's objective function to penalize biases towards *popular items*. For instance, regularization terms can be added to the objective function to penalize recommendations that skew too heavily towards *popular items*. Alternatively, models can be designed to explicitly predict the popularity of items and adjust

recommendations accordingly. Machine learning techniques, such as adversarial training, have also been applied to create models that are inherently less sensitive to the *popularity bias* present in the training data. These methods seek to directly address the root of the bias within the recommendation process itself, promoting a more equitable distribution of item recommendations.

Post-processing methods adjust the output of the recommendation algorithm to counteract *popularity bias*. One approach is to *re-rank* the list of recommended items to ensure that less popular items are more prominently featured. This can be achieved by applying a diversity-promoting *re-ranking* algorithm that explicitly takes item popularity into account, ensuring that the final list of recommendations includes a mix of popular and *niche* items. Another strategy involves applying thresholding techniques to filter out excessively popular items from the recommendation list, thereby giving more visibility to less popular items. These methods allow for the preservation of the original recommendation algorithm while still addressing the issue of *popularity bias* in the final recommendations.

Debiasing methods in recommender systems are critical for ensuring that recommendations reflect a wide variety of user interests and promote content diversity. By implementing *pre-processing*, *in-processing*, and *post-processing* techniques, researchers and practitioners can mitigate the effects of *popularity bias*, thereby enhancing user satisfaction and engagement. These methods represent a step towards more equitable and diverse recommender systems, where all items, regardless of their inherent popularity, have a fair chance of being discovered and appreciated by users. As the field continues to evolve, further research and experimentation will be essential in developing more sophisticated debiasing techniques that can adapt to the complexities of user preferences and the dynamic nature of content popularity.

Post-Processing Debiasing Approach

Post-processing methods in recommender systems serve as a crucial strategy to mitigate popularity bias after the generation of initial recommendations. These techniques intervene at the final stage, adjusting the recommended item list to ensure a more balanced exposure of items across different popularity levels. The core aim is to refine the recommendations by either demoting overly popular items or promoting *long-tail* (*less popular*)

items, thereby addressing the skewness towards popular items that many recommendation algorithms exhibit. This section delves into various *post-processing* strategies, highlighting their principles, methodologies, and implications for reducing popularity bias, as discussed in kind of studies.

Re-ranking is one of the most common *post-processing* approaches to mitigate popularity bias. It involves adjusting the order of items in the recommendation list based on criteria other than the original recommendation scores. This can include incorporating diversity or novelty metrics that explicitly account for item popularity. For instance, a simple yet effective re-ranking strategy might involve penalizing the scores of popular items or boosting the scores of less popular items, thereby altering the final recommendation list to include a more representative mix of items.

An advanced *re-ranking* method, as detailed in several studies, employs more sophisticated criteria that consider the user's historical preferences for popular versus *niche* items. By tailoring the *re-ranking* process to individual user profiles, these methods aim to achieve a personalized balance between relevance and diversity, enhancing user satisfaction by exposing them to a broader range of content.

Diversity enhancement techniques specifically focus on increasing the *heterogeneity* of the recommended items to counteract the dominance of popular items. This involves selecting items that not only match the user's interests but also contribute to a diverse recommendation list. One approach to *diversity enhancement* is to maximize the dissimilarity between recommended items, ensuring that the final list covers a wide range of item categories or genres.

Incorporating *diversity metrics* into the *post-processing* stage allows for a systematic adjustment of recommendations to achieve desired diversity levels. Studies has proposed various metrics for quantifying diversity, such as intra-list diversity, which measures the dissimilarity among items within a recommendation list. By optimizing these *diversity metrics*, recommender systems can deliver more varied and less biased recommendations.

Fairness-oriented post-processing methods aim to ensure equitable exposure for items, particularly focusing on promoting *fairness* among items with different popularity levels. These methods often involve adjusting the recommendation algorithm's output to achieve a more balanced representation of both *popular* and *niche* items, thereby ensuring that all

items have a fair chance of being recommended.

One approach to promoting *fairness* is to implement exposure control mechanisms that allocate a certain portion of the recommendation list to less popular items. This ensures that these items receive a minimum level of exposure, regardless of their inherent popularity. Such methods not only help in reducing popularity bias but also in enhancing the *long-tail* discovery, allowing users to explore a wider range of content.

Post-processing methods play a pivotal role in mitigating popularity bias by adjusting the recommendations after they have been generated. Through *re-ranking*, *diversity enhancement*, and *fairness-oriented strategies*, these methods strive to correct the biases inherent in the initial recommendation lists. By promoting a more balanced exposure of items, *post-processing* techniques contribute to improving the overall quality and fairness of recommendations. As recommender systems continue to evolve, the development and refinement of *post-processing* methods will be crucial in addressing the challenges posed by popularity bias, ensuring that recommendations serve the interests of both users and content providers equitably.

In-Processing Debiasing Approach

In-processing methods for mitigating *popularity bias* in recommender systems involve incorporating adjustments directly into the recommendation algorithms to reduce the influence of popular items in the recommendation process. These techniques are designed to alter the core mechanism of recommendation models, ensuring that the recommendations generated reflect a more balanced distribution of items across different popularity levels. The objective of *in-processing* methods is to integrate bias mitigation into the model's learning process, thereby inherently reducing the system's tendency to favor popular items over less popular ones. This section delves into several *in-processing* strategies, including *regularization approaches*, *bias-aware models*, and *adversarial training*, highlighting their methodologies and impacts on addressing popularity bias.

Regularization techniques are a cornerstone of *in-processing* methods for debiasing. They involve modifying the recommendation model's objective function by adding a penalty term designed to reduce the model's propensity to favor popular items. For instance, a regularization term can be added to penalize the model when it excessively recommends

popular items, thereby encouraging the model to explore and recommend a broader range of less popular items. This approach can be seen as a way to explicitly incorporate the diversity of recommendations into the model optimization process.

One popular *regularization technique* is the incorporation of a diversity regularization term that measures the concentration of recommendations towards popular items. By penalizing high concentration, the model is forced to consider less popular items, thus reducing the overall popularity bias. The effectiveness of *regularization-based* methods is contingent on the careful tuning of *hyperparameters* that control the strength of the bias mitigation versus the accuracy of the recommendations.

Bias-aware recommendation models explicitly account for the popularity of items as part of the recommendation process. These models are designed to understand and adjust for the inherent bias towards popular items present in the user-item interaction data. One approach involves creating models that have separate components for predicting user-item interactions and for modeling the popularity distribution of items. By doing so, the recommendation system can differentiate between genuine user preferences and the effects of popularity bias.

Another *bias-aware* method involves the use of *factorization models* that incorporate item popularity as an *explicit* factor in the prediction process. These models adjust their recommendations based on an item's popularity level, ensuring that items are recommended based on user preferences rather than their widespread popularity. This nuanced approach allows for a more personalized recommendation experience that accounts for the diversity of user interests.

Adversarial training introduces a game-theoretic framework to the recommendation process, where the recommendation model (the generator) is trained alongside an adversary that attempts to identify recommendations influenced by popularity bias. The objective of the recommendation model is to generate recommendations that the adversary cannot distinguish from a distribution of recommendations that reflects no popularity bias. This process encourages the recommendation model to learn representations and patterns that are not solely based on item popularity, thereby reducing the bias in its recommendations.

The adversarial training process involves alternating between updating the recom-

mendation model to minimize the traditional loss function while also minimizing the ability of the adversary to detect bias, and updating the adversary to maximize its ability to detect biased recommendations. This technique has shown promise in creating recommendation models that are more *robust* to popularity bias, as it forces the model to explore and recommend items that it would otherwise overlook due to their lower popularity.

In-processing methods offer a powerful approach to mitigating popularity bias in recommender systems by embedding bias reduction mechanisms directly into the *recommendation algorithms*. Whether through *regularization*, *bias-aware modeling*, or *adversarial training*, these techniques aim to ensure that the recommendations generated serve a broad spectrum of user interests and promote content diversity. By addressing popularity bias at the model level, *in-processing* methods provide a pathway towards more equitable and personalized recommendation systems that are capable of surfacing a wider variety of items to users. As research in this area continues to evolve, it is expected that new and more sophisticated *in-processing* methods will emerge, further enhancing the ability of recommender systems to offer diverse and unbiased recommendations.

Pre-Processing Debiasing Approach

Pre-processing methods in recommender systems aim to address and *mitigate popularity bias* before the data is utilized by the recommendation algorithms. These techniques focus on adjusting the input data to reduce the inherent biases present, ensuring a more balanced representation of items across different popularity levels. The objective is to *pre-process* the user-item interaction data in such a way that the recommendation models trained on this data are less likely to perpetuate or exacerbate *popularity bias*. This section explores various *pre-processing* strategies, including re-sampling, data augmentation, and bias correction techniques, drawing from researches to illustrate their methodologies and impacts on reducing *popularity bias*.

Re-sampling involves modifying the distribution of the user-item interaction data to counteract the over-representation of *popular items*. This can be achieved by either under-sampling interactions with *popular items* or oversampling interactions with *less popular* items. The goal is to create a more balanced dataset that represents a wider array of items, thereby reducing the model's tendency to favor *popular items* during the recommendation

process.

One notable approach within *re-sampling* is the implementation of a *popularity-based* weighting scheme, where each user-item interaction is weighted inversely proportional to the item's popularity. This method, discussed in various studies, ensures that interactions with *less popular* items contribute more significantly to the model's learning process, promoting a more equitable representation of items. The effectiveness of *re-sampling* methods largely depends on the careful calibration of sampling rates to achieve the desired balance without introducing new biases or significantly compromising the recommendation quality.

Data augmentation in the context of recommender systems involves artificially augmenting the dataset with additional user-item interactions, particularly focusing on enhancing the representation of *less popular* items. This can be accomplished through techniques such as synthetic interaction generation, where new interactions are created based on the patterns observed in the existing data, but with a deliberate focus on *less popular* items.

Another form of *data augmentation* is the extrapolation of existing interactions to similar items, particularly those that are *less popular*. This approach is grounded in the assumption that if a user likes a certain item, they may also like similar items, even if those items are not among the *most popular*. By augmenting the data with these inferred interactions, recommender systems can achieve a more balanced coverage of the item space, thereby reducing *popularity bias*.

Bias correction techniques involve directly modifying the user-item interaction data to mitigate the effects of *popularity bias*. This could include adjusting the interaction values to diminish the impact of frequent interactions with popular items or incorporating *popularity bias* correction factors into the data *pre-processing* stage.

One sophisticated approach to *bias correction* is the use of statistical models to estimate the extent of *popularity bias* within the dataset and then apply corrections to the interaction data to counterbalance this bias. This method requires a thorough understanding of the data's underlying distribution and a careful calibration of correction parameters to ensure that the bias is effectively mitigated without distorting the data's integrity.

Pre-processing methods play a crucial role in tackling *popularity bias* in recommen-

der systems by addressing the bias at its source: *the input data*. By employing strategies such as *re-sampling*, *data augmentation*, and *bias correction*, these techniques strive to create a more equitable data foundation upon which recommendation models can be built. The ultimate goal is to produce recommendation systems that are capable of surfacing a diverse range of items, thereby enhancing user satisfaction and promoting content discovery. As the field advances, further research and development of *pre-processing* methods are essential to overcoming the challenges posed by *popularity bias* and achieving more balanced and fair recommendations.

In this study, organize this approach against four distinguished strategies designed to counteract popularity bias: *Explicit Query Aspect Diversification (xQuad)* (Abdollahpouri et al. 2019), *Augmentative (AUG)* (Yalcin and Bilge 2021), *Popularity-aware Weighting (PAW)* (Abdollahpouri et al. 2018), *Multiplicative (MUL)* (Yalcin and Bilge 2021), and *Largest-Normalized-Score-First (LNSF)* (Dong et al. 2021). The subsequent sections provide a comprehensive description of these methodologies:

1. *Explicit Query Aspect Diversification (xQuad)* (Abdollahpouri et al. 2019): This method, recognized for its effectiveness in mitigating popularity bias, categorizes items into two groups: popular (i.e., *head*) and less popular (i.e., *tail*). It strives to ensure a balanced representation of these groups in the final recommendations, taking into account user preferences. The balance between accuracy and diversity is regulated by a control parameter, λ , which is typically set to 0.5.
2. *Augmentative (AUG)* (Yalcin and Bilge 2021): The *AUG* strategy merges actual prediction scores with the inverse weights of item popularity to generate synthetic ranking scores. Here, the prediction scores act as the main driver, while the popularity weights function as an additive factor, prioritizing ranking accuracy over the reduction of popularity bias.
3. *Popularity-aware Weighting (PAW)* (Abdollahpouri et al. 2018): This technique employs logarithmic inverse weights for items based on their popularity, allocating increased weight to items of lesser popularity and diminishing weight to those more popular. These weights are then combined with prediction scores to compute

ranking scores for the ultimate recommendations. A parameter λ , adjusted to 0.5, is utilized to balance these dual aspects of recommendations.

4. *Multiplicative (MUL)* (Yalcin and Bilge 2021): While sharing the *re-ranking strategy* with the *AUG* that penalizes popular items, *MUL* adopts a distinctive approach. It incorporates item weights as a multiplicative factor in the computation of final ranking scores, thereby significantly favoring less popular items in the recommendations.
5. *Largest-Normalized-Score-First (LNSF)* (Dong et al. 2021): This method is aimed at promoting exposure *fairness* and mitigating popularity bias. It implements stock volume constraints grounded on historical interaction data to limit recommendations, thus diminishing the Matthew Effect concerning item popularity. The *Minimum Cost Maximum Flow* model is utilized to optimally align users with items under these restrictions.

2.5. Datasets

Recommender systems are pivotal in guiding users through a vast array of choices in various domains such as e-commerce, content streaming, and social networking. The effectiveness of these systems largely depends on the quality and characteristics of the datasets they utilize. This study explores three significant datasets widely used in the recommender systems domain: *MovieLens 1M*, *DoubanBook*, and *Yelp Verified*. Each dataset's unique features and their implications for recommender system development and evaluation are discussed.

MovieLens-1M dataset, created by the *GroupLens Research Group*, serves as one of the cornerstone datasets for movie recommendation system research. Comprising approximately one million ratings from over 6,000 users on around 3,900 movies, this dataset offers a rich foundation for developing and testing recommendation algorithms. Ratings are on a scale from 1 to 5, and the dataset includes user demographic information (age, gender, occupation, and ZIP code) as well as detailed movie metadata (genres, title, release year, etc.). The diversity of user profiles and the comprehensive metadata for movies make the *MovieLens 1M* dataset an ideal

playground for testing various recommender system algorithms. It supports not only *collaborative filtering* approaches but also *content-based* and *hybrid models*, providing a versatile basis for algorithm development and comparative studies. The inclusion of user demographic information further enriches the dataset, allowing for the exploration of personalized recommendation strategies that consider user-specific characteristics. While the *MovieLens 1M* dataset is extensively used due to its *robust* size and detailed information, it also presents challenges such as the *sparsity* of the user-item interaction matrix and potential biases in user ratings. These challenges prompt researchers to develop more sophisticated algorithms that can handle sparse data and *mitigate biases*, pushing the boundaries of recommender system capabilities.

DoubanBook is a dataset derived from *Douban*, a leading *Chinese social networking service* that offers book recommendations among other content. This dataset encapsulates book ratings and rich metadata, including author, publisher, publication year, and ISBN, collected from millions of users and hundreds of thousands of books. The extensive user interaction data, including ratings, reviews, and social networking features, offers a unique dataset for book recommendation systems. *DoubanBook* is particularly valuable for developing and evaluating book recommendation algorithms. It allows researchers to explore *content-based filtering*, *collaborative filtering*, and *hybrid approaches* by leveraging the rich book metadata and user interactions. The dataset's *diversity* in user preferences and book genres provides a comprehensive testbed for examining the effectiveness of different recommendation strategies and understanding user reading habits. The complexity of user preferences and the dynamic nature of book popularity pose significant challenges in using the *DoubanBook* dataset effectively. Researchers are tasked with devising algorithms that can accurately predict user preferences while adapting to trends and new book releases. Additionally, the social networking aspect of *Douban* introduces another layer of complexity, enabling studies on the impact of social influence on book recommendations.

Yelp Verified dataset is an extensive collection of verified user reviews and ratings

for local businesses, primarily restaurants and cafes, gathered from the *Yelp* platform. It includes detailed information on business categories, locations, user demographics, and the textual content of reviews. This dataset is particularly useful for developing recommender systems aimed at local services, incorporating user feedback and sentiment analysis. *Yelp Verified* offers a valuable resource for understanding consumer preferences and behaviors regarding local businesses. The dataset supports the development of location-based recommendation systems and allows for the exploration of how user sentiments expressed in reviews can influence recommendation quality. The granularity of the dataset facilitates detailed analyses of user-business interactions, enabling personalized and context-aware recommendations. Utilizing the *Yelp Verified* dataset involves navigating challenges such as the variability in review quality, the need for effective sentiment analysis tools, and handling the geospatial dimensions of the data. These challenges provide opportunities for research into advanced natural language processing techniques and the integration of geographical information systems (GIS) into recommender systems. The potential for cross-domain recommendations and the exploration of new recommendation paradigms makes *Yelp Verified* a rich dataset for innovative recommender system research.

In this experiments, employ three publicly accessible real-world benchmark datasets gathered from distinct application domains, as highlighted in a closely related recent study (Yalcin and Bilge 2022); specifically, MovieLens-1M⁷ (ML) for movies (Harper and Konstan 2015), Douban Book⁸ (DB) for books (Shi et al. 2018), and Yelp⁹ for local business reviews (Shi et al. 2018). User ratings across all datasets are given in a discrete five-star scale. Detailed characteristics of the ML, DB, and Yelp datasets are provided in Table 2.3. These datasets vary in dimensions and sparsity levels, as detailed in Table 2.3, allowing us to examine the relationship between these critical dataset attributes and the potential for unfairness among user personas with varying levels of privacy.

⁷<https://grouplens.org/datasets/movielens/1m/>

⁸<https://www.douban.com/>

⁹<https://www.yelp.com/>

Table 2.3. Characteristics of the MovieLens-1M (ML), Douban Book (DB), and Yelp datasets

Dataset	# of Users	# of Items	# of Ratings	Sparsity (%)	Rating Average	# of Ratings per Item	# of Ratings per User
ML	6,040	3,952	1,000,209	95.8	3.58	253.1	165.6
DB	13,024	22,347	792,062	99.7	4.05	35.4	60.8
Yelp	16,239	14,284	198,397	99.9	3.77	13.9	12.2

2.6. Evaluation Methods

Recommender systems have become integral to the digital experience, guiding users in a myriad of contexts, from e-commerce to content streaming. The essence of these systems lies in their ability to predict and recommend items that are most likely to be of interest to the user, thereby enhancing user experience and engagement. Given the critical role of recommender systems, evaluating their performance accurately is paramount. This work delves into the characteristics of recommender systems, the importance of evaluation, and the methodologies employed to assess their performance.

Recommender systems are characterized by their ability to process large volumes of data to identify patterns, preferences, and potential interests of users. They can be categorized into three primary types based on their filtering approaches: *content-based*, *collaborative filtering*, and *hybrid methods*. *Content-based systems* recommend items similar to those a user has liked in the past, based on item features. *Collaborative filtering* systems predict user preferences based on the likes and dislikes of similar users. *Hybrid systems* combine both approaches to leverage their strengths and mitigate their weaknesses. Evaluation models also provide metrics for all sub-approaches and system performance can be monitored with various metrics and points for improvement can be identified.

The evaluation of recommender systems is critical for several reasons. Firstly, it allows the determination of a system's effectiveness in meeting user needs and preferences. Secondly, it aids in comparing the performance of different recommendation strategies and algorithms, guiding the selection or development of superior models. Lastly, evaluation provides insights into areas of improvement, driving the refinement of recommender systems to enhance accuracy, relevance, and user satisfaction.

Evaluation of recommender systems includes both *offline* and *online* methodologies. However, this study was conducted only on offline measurements. To elaborate briefly;

- *Offline Evaluation*: Utilizes historical data to assess the performance of recommendation algorithms without the need for real-time user interaction. It is efficient and widely used for preliminary evaluation.
- *Online Evaluation*: Involves real-time user interaction, typically through A/B testing, to measure actual user response to recommendations. Online evaluation provides insights into user satisfaction, engagement, and behavioral patterns that offline methods cannot capture.

Evaluating the performance of recommender systems involves various metrics and methodologies, each suited to different aspects of performance. The choice of evaluation metrics is influenced by the specific objectives of the recommender system, including *accuracy*, *diversity*, *novelty*, *coverage* and *entropy* etc.

2.6.1. Accuracy Metrics

Accuracy is one of the most fundamental criteria for assessing recommender systems. It measures how closely the recommendations match the actual interests of the users. Common accuracy metrics include:

Precision, Recall and F1-Score: For assessing the accuracy of the top- N recommendation lists generated, utilize the recognized metrics of *Precision*, *Recall*, and *F1-score* (Yalcin and Bilge 2022). Specifically, the *Precision* rate (i.e., $P@N_u$) for a user u is defined as the proportion of recommended items within the top- N list that are deemed suitable for the user. Conversely, the *Recall* rate (i.e., $R@N_u$) for the same list is the fraction of the total suitable items recommended out of all items rated in the user u 's profile. In determining the suitability of an item, a threshold of 3.5 is adopted, aligning with the consensus in literature that rates 4 and 5 on a 5-star scale as positive (Bobadilla et al. 2013). Lastly, the *F1-score* (i.e., $F1@N_u$) for the list is derived as the harmonic mean of the $P@N_u$ and $R@N_u$ rates, as depicted in Eq. 2.1, thereby offering a balanced evaluation between these two accuracy metrics

from distinct perspectives. The harmonic mean of precision and recall, providing a single metric to balance the two in F1-Score.

$$F1@N_u = 2 \times \frac{P@N_u \times R@N_u}{P@N_u + R@N_u} \quad (2.1)$$

2.6.2. Beyond Accuracy Metrics

This part provides a comprehensive explanation of both the accuracy (namely, $nDCG$) and the beyond-accuracy metrics (specifically, $APLT$, $Entropy$, $Novelty$, and LTC) used in evaluation.

Average Popularity of Recommended Items (APRI): This metric assesses the average popularity level of items recommended by the algorithms. A higher $APRI$ score in top- N lists recommends that the recommendation system tends to favor items with a higher degree of popularity bias, which is undesirable (Yalcin 2022b). In detail, it initially determines the popularity ratio P_i for each recommended item i , based on the proportion of total users who have rated the item against the total user base. Subsequently, it calculates the mean of these P values for items within the Top- N selection, resulting in an $APRI_u$ value for the user u , as outlined in Eq. 2.2.

$$APRI_u = \frac{\sum_{i \in N_u} P_i}{N} \quad (2.2)$$

Normalized Discounted Cumulative Gain (nDCG): The $nDCG$ metric is pivotal in this research as a statistical indicator of accuracy. It evaluates the genuine ratings awarded to items and their corresponding positions within the top- N recommendations. The true rating that a user u assigns to an item i is represented as $r_{u,i}$. Leveraging this, calculate the $DCG_{N_u}^u$ and $nDCG_{N_u}^u$ for the top- N list of user u by adhering to the equations presented in Eqs. 2.3 and 2.4.

$$DCG_{N_u}^u = r_{u,i_1} + \sum_{n=2}^{N_u} \frac{r_{u,i_n}}{\log_2(n)} \quad (2.3)$$

$$nDCG_{N_u}^u = \frac{DCG_{N_u}^u}{IDCG_{N_u}^u} \quad (2.4)$$

Within this framework, the $IDCG_{N_u}^u$ signifies the maximum achievable gain for user u , realized by organizing N items in the ideal order.

Average Percentage of Long Tail Items (APLT): In addition to the accuracy metrics discussed earlier, employ the *Average Percentage of Long Tail Items (APLT)* for evaluating ranking-based recommendations. The *APLT* metric quantifies the share of *long-tail* items in the recommendation list, thereby indicating the algorithm's ability to recommend more *niche* and uncommon items. This metric is crucial for gauging how effectively a debiasing approach can promote less popular (i.e., *tail*) items within its recommendations. The selection of *tail* items from the entire item pool is guided by the *Pareto Principle* (Sanders 1987). With T representing the set of designated *tail* items, the *APLT* score for a user u 's top- N list is calculated as per the equation given in Eq. 2.5.

$$APLT_u = \frac{|i \in T \cap N_u|}{N} \quad (2.5)$$

It's critical to acknowledge that while the $APLT_u$ score reflects the recommendation diversity for an individual user, the overall *APLT* score essentially represents the average of these diversification levels across all users. Therefore, a high *APLT* score does not necessarily indicate that the algorithm uniformly provides diverse and high-quality recommendations. Conversely, a low *APLT* score explicitly indicates a widespread lack of diversity within the recommendation lists. For this reason, additionally employ *Entropy* as a complementary metric to assess the diversity across all recommended lists, which is discussed in more detail below.

Novelty: In this study, employ the *Novelty* metric as a supplementary gauge that extends beyond simple accuracy measures. This metric evaluates the ability of recommendation systems to propose items previously unrated by a user. Specifically, *Novelty* measures the ratio of unencountered items within the top- N recommendation list for a user, showcasing the algorithm's potential to offer novel selections. To compute the *Novelty* score for an individual user u , first define I_u as the set of

items already rated by user u . The *Novelty* score for the top- N recommendations presented to user u , denoted as $Novelty_u$, is then determined based on the approach outlined in Eq. 2.6

$$Novelty_u = \frac{|i, i \in N_u \text{ and } i \cap I_u = \emptyset|}{N} \quad (2.6)$$

In this setting, N_u represents the set of top- N items recommended to user u .

Long-Tail Coverage (LTC): In addition to the accuracy metrics previously discussed, this study further evaluates the *Long-Tail Coverage (LTC)* metric to assess the representation of lesser-known items in the catalog, as delineated in (Abdollahpour 2020). The *LTC* metric is computed by aggregating the top- N recommendation lists from each user into a singular set denoted as $\mathbb{N}$. This aggregated set $\mathbb{N}$ excludes any duplicates found across individual top- N lists. Subsequently, determine the intersection of items, $I_{\mathbb{N} \cap T}$, that are found both in $\mathbb{N}$ and the designated *tail item* set T , following the *Pareto Principle*. The efficacy of a recommendation algorithm's *LTC* score is quantified by the ratio of $I_{\mathbb{N} \cap T}$ to the overall count of T , as elaborated in Eq. 2.7. An elevated *LTC* score indicates the algorithm's superior capability in incorporating a diverse array of less dominant, or tail, items within its recommendations.

$$LTC = \frac{|I_{\mathbb{N} \cap T}|}{|T|} \quad (2.7)$$

Entropy: As investigated by (Elahi et al. 2021), the *Entropy* metric assesses how frequently different items are recommended across the entire catalog by an algorithm. Essentially, it quantifies the variability in the recommendation frequency of each item. An algorithm that scores higher on *Entropy* suggests a more equitable distribution of recommendations across items, enhancing diversity. Such uniformity is vital for fostering a competitive and equitable market environment. To compute the *Entropy* score, one first consolidates the top- N recommendation lists from all users, accounting for repeated item appearances. Let this aggregated set be denoted by N_{all} , and let $i_1, i_2, \dots, i_K$ represent the complete set of items in the catalog,

where K is the total number of items. The procedure for calculating the *Entropy* of a recommendation algorithm is described in Eq. 2.8.

$$Entropy = - \sum_{i \in K} (\Pr(i)) \log_2(\Pr(i)) \quad (2.8)$$

Here, $\Pr(i)$ denotes the relative frequency of the appearance of item i within the N_{all} set.

The evaluation of recommender systems is a multifaceted process, essential for optimizing system performance and enhancing user satisfaction. By employing a combination of accuracy metrics and considering factors like *diversity*, *novelty*, and *coverage*, developers can gain a comprehensive understanding of a system's efficacy. Both *offline* and *online* methodologies play crucial roles in this evaluation process, each offering unique insights into the recommender system's impact on user experience. As recommender systems continue to evolve, so too will the strategies for their evaluation, ensuring that they remain aligned with the ever-changing landscape of user needs and preferences.

3. LITERATURE REVIEW

Given that this study primarily focuses on two critical issues within recommendation systems, namely *privacy preservation* and the challenge of *popularity bias*, have divided this section into two parts. The first part conducts a review of the literature on strategies for preserving privacy specifically within *collaborative filtering (CF)* recommendation systems. The second part, in contrast, explores recent research addressing the issue of *popularity bias*, including various approaches devised to mitigate this problem.

3.1. Privacy-Preserving Approaches for CF Recommenders

In recent times, there has been a notable surge in research addressing the privacy issues associated with traditional *collaborative filtering (CF)* systems. Numerous studies have been devoted to creating *privacy-preserving collaborative filtering (PPCF)* strategies to tackle these concerns (Wei et al. 2018). Typically, these strategies employ methods such as *k*-anonymity, obfuscation, differential privacy, cryptography, and *perturbation-based techniques* to ensure user privacy.

In the realm of distributed recommendation systems, cryptographic methods are commonly employed to ensure the privacy protection of users (Badsha et al. 2016; Li et al. 2017). For instance, a novel unsynchronized secure multi-party computation protocol was proposed by Li et al., allowing for the incremental calculation of item similarities without necessitating the presence of individuals during the computation phases (Li et al. 2016). Additionally, introduced a distributed recommendation mechanism where merchants exchange encrypted information through secure channels with an intermediary to perform predictions or to rank items for recommendation lists (Shmueli and Tassa 2017). Moreover, recent research has utilized homomorphic encryption to secure Quality of Service metrics for recommending personalized web services tailored to user and location data, thereby safeguarding sensitive information (Badsha et al. 2018). Some studies have also put forward an effective strategy for group recommendation systems designed to protect the privacy of users during the recommendation process (Li et al. 2017). According to Zou et al., calculating item similarities is treated as a probabilistic inference problem, with a semi-distributed belief propagation network method being introduced for item-based

PPCF systems, effectively preventing the exposure of user preferences by only involving a specific subset of users in the network at any given time (Zou and Fekri 2015).

Anonymization stands as a key technique in user *privacy protection*, focusing on severing the identifiable link between individuals and their rating data. Casino advocate for a micro aggregation-based *PPCF* approach that employs k -anonymity masks to safeguard user privacy (Casino et al. 2015). Chen delve into achieving k -anonymity for voluminous and sparse databases (for instance, Recommendation Systems) and present a clustering-based k -anonymity heuristic method aimed at privacy preservation in data exchange (Chen and Huang 2012). Moreover, Wei introduce a (p, l, a) -diversification strategy to enhance the efficacy of established k -anonymity techniques, where p represents prior knowledge about users' rating profiles and (l, a) symbolize the diversification of users to augment the level of *privacy protection* (Wei et al. 2018). As a final instance of anonymization-centric methods, Zhang implement this tactic within the context of privacy-conscious smart city environments, conceptualizing urban privacy dilemmas as a *PPCF* challenge (Zhang et al. 2018).

Differential privacy serves as a key mechanism in CF methodologies to measure the potential privacy compromise when incorporating an individual's preferences into recommendations. The seminal work by Dwork introduced the concept of differential privacy, which was further applied in the realm of recommendation systems by McSherry et al. through the randomization of actual user preferences via a proprietary covariance matrix (Dwork et al. 2006; McSherry and Mironov 2009). A novel approach by Guerraoui et al. introduces a distance-based differential *privacy* method, modifying real user profiles by swapping elements at certain distances within these profiles (Guerraoui et al. 2015). Additionally, Shen recommends a unique approach that integrates noise calibration to provide theoretical *privacy* protection and utility, facilitating the injection of instance-specific noise into users' preference data (Shen and Jin 2014). Lastly, Hou develop a dual-phase strategy for implementing neighbourhood-based differential privacy in recommendations and secure neighbour selection, specifically tailored for the medical recommendation scenario (Hou et al. 2018b).

In *collaborative filtering* (CF) frameworks, strategies for altering data such as *obfuscation* and random *perturbation* stand as the primary means for safeguarding *privacy*, as

identified in scholarly works. Specifically, *obfuscation* strategies aim to obscure personal details while preserving the utility of the data for recommendation purposes. For instance, a permutation-based *obfuscation* approach for CF systems operating on central servers is introduced by Parameswaran and Blough (2007). Moreover, Badsha et al. describe a user-centric *PPCF* model that employs semi-honest intermediaries for conducting ancillary computations (Badsha et al. 2017). In scenarios involving distributed recommendations, Boutet et al. advocate for the use of *obfuscated* user profiles with coarse granularity to protect sensitive information (Boutet et al. 2016). Furthermore, Elmisery advance the methodology by applying varying degrees of *obfuscation* to enhance the identification of trustworthy target users beyond the conventional single-level *obfuscation* approaches (Elmisery and Botvich 2017).

Randomized Perturbation Techniques (RPTs) serve as an innovative alternative to traditional methods, offering protection for user privacy by altering genuine user preferences prior to their transmission to the recommendation server (Polat and Du 2003; Polat and Du 2005a; Polat and Du 2005b). As an illustration, Bilge have engineered a scalable *privacy-preserving collaborative filtering (PPCF)* methodology that leverages bisecting *k*-means clustering along with randomized perturbations (Bilge and Polat 2013). Furthermore, Gong ensure the confidentiality of user profiles dispersed across different storage systems in the recommendation pipeline by merging *RPTs* with secure multi-party computation strategies (Gong 2011). To augment the efficacy of the PPCF framework based on *RPTs*, Polar et al. advocate for the application of a diverse spectrum of random values (Polatidis et al. 2017). Moreover, Liu have formulated a hybrid model that amalgamates *RPTs* with the principles of differential privacy, thereby providing a more fortified safeguard against breaches of user preferences than existing RPT-centric models (Liu et al. 2017). In their contribution to the field, Yargic have successfully applied RPTs within the realm of multi-criteria recommendation systems to ensure the privacy of individual users (Yargic and Bilge 2019).

Beyond the realm of Recommender Systems (RSs), there are notable advancements in *privacy preservation* mechanisms, particularly within the field of machine learning and more specifically, in federated machine learning (Yin et al. 2021). For instance, Xu have unveiled *HybridAlpha*, a novel system designed to mitigate privacy threats in federated

learning environments through the use of Secure Multiparty Computation (SMC) protocols derived from functional encryption techniques (Xu et al. 2019). In another innovative study, a hybrid method that merges SMC with differential privacy has been proposed to navigate the delicate balance between maintaining accuracy and managing small datasets within participating entities (Truex et al. 2019). Furthermore, Kerkouche have developed a privacy enhancement strategy for federated learning specifically tailored for the healthcare sector, utilizing differential privacy to boost bandwidth efficiency while securing confidential data (Kerkouche et al. 2021). Lastly, Lu have proposed a federated learning strategy aimed at identifying potential failures in wind turbines within an energy organization (Lu et al. 2022). To safeguard cross-organizational data privacy, this approach incorporates a biometric authentication system, significantly diminishing the risk to privacy.

Recommender Systems (RSs) are known to rely on individual rating data to generate suitable recommendations for users. Consequently, the volume or dimensionality of data sets used for prediction significantly impacts the quality of recommendations. This demand for large datasets introduces two main challenges. The first challenge involves the costs associated with transferring, analyzing, and processing such extensive data. The second challenge is the difficulty in ensuring the confidentiality of large datasets and safeguarding user privacy. To address these issues, recent research has focused on leveraging distributed environments to enhance data privacy (Omar et al. 2019; Bosri et al. 2020; Lin et al. 2021). For instance, Bosri have unveiled *Private-Rec*, a system that marries artificial intelligence techniques with blockchain technology to create a privacy-centric recommendation system (Bosri et al. 2020). This system ensures a secure setting for users by leveraging distributed ledger technology, which processes users' data only with their consent. Another study has presented a blockchain-enabled recommender system that emphasizes privacy protection (Omar et al. 2019). The decentralized framework of blockchain empowers users to have sovereignty over their data and agree to its utilization in analyses. This model motivates users to contribute their data by providing incentives like rewards, points, or discounts, applicable to the offerings of the online platform. Lastly, Lin have developed a blockchain-integrated recommendation system that employs local sensitive hashing and differential privacy within a blockchain context, effectively reducing the computational

costs of predictions while simultaneously enhancing both recommendation accuracy and privacy protection for users (Lin et al. 2021).

3.2. Popularity Bias Issue of CF Recommenders

A paramount issue within Recommender Systems (RSs) pertains to the fairness of the system towards its two main stakeholders: the users and the providers of items or services, ensuring that the recommendations do not perpetuate any form of bias (Abdollahpouri and Burke 2019; Yalcin 2022a). Nevertheless, recent investigations have identified that recommendation algorithms often struggle to equitably manage the items within their generated rankings, attributable to various factors including the inherent algorithms' mechanisms or the characteristics of the datasets on which they are trained (Chen et al. 2020).

A prime example of this challenge is the issue of *popularity bias*, which results in recommendation lists being predominantly filled with popular items, regardless of their actual desirability (Abdollahpouri et al. 2020; Yalcin and Bilge 2022). Unfortunately, this leads to a marked reduction in the quality of recommendations beyond mere *accuracy*, affecting aspects such as *diversity*, *catalog coverage*, and *novelty*, and further compromises the system's resilience to certain types of malicious attacks. Consequently, there's an increasing focus on examining the detrimental impacts of this bias and on devising effective strategies for *mitigating popularity bias*, in order to generate more balanced and suitable recommendations (Abdollahpouri et al. 2019; Yalcin and Bilge 2021; Boratto et al. 2021).

Initial investigations into this issue primarily uncovered its impact across various application areas such as movies, music, travel, and online courses. For instance, Jannach examined multiple *collaborative filtering (CF)* recommendation systems across different methodologies, including *matrix factorization* and *neighborhood-based approaches*, and identified a tendency to repeatedly recommend a limited set of popular items (Jannach et al. 2015). Furthermore, they explored how the fine-tuning of parameters within these CF algorithms influences their predisposition towards popular items. In a related vein, Boratto conducted an analysis on massive open online courses, utilizing the same CF recommendation systems to determine their propensity for bias towards the popularity of certain courses and their categories (Boratto et al. 2019). Another groundbreaking study

by Abdollahpouri delves into the varying impacts of this issue on both consumers and suppliers of items (Abdollahpouri et al. 2020).

The body of research also explores the impact of recommenders' *popularity bias* on users with varying levels of interest in item popularity, noting how it disproportionately disadvantages certain groups. For instance, in the context of movie recommendations, Abdollahpouri et al. found that users with a preference for popular items suffer more from *popularity bias* compared to those who favor *niche* items (Abdollahpouri et al. 2019). This analysis has been extended to music recommendations by examining a broader range of *collaborative filtering (CF)* recommenders (Kowald et al. 2020). Additionally, Yalcin and Bilge segmented users based on key characteristics of their rating behaviors, revealing that users who are highly interactive, selective, and difficult to predict face particularly unfair recommendations, especially when considering their original inclination towards item popularity (Yalcin and Bilge 2022). Furthermore, Yalcin and Bilge investigated the influence of popularity bias on individuals possessing different personality traits as defined by the big-five-factor model, finding that those averse to new experiences and less extroverted individuals receive more biased recommendations in terms of item popularity (Yalcin and Bilge 2023).

Scholars have been investigating strategies to diminish the impact of *popularity bias* within recommender systems. In pursuit of this goal, a variety of techniques have been introduced, which can generally be segmented into three distinct approaches: *pre-processing*, *in-processing*, and *post-processing*, each distinguished by their method of bias reduction in the recommendation process (Boratto et al. 2021). *Pre-processing* techniques focus on adjusting the dataset before training recommendation models to ensure a more balanced representation of user interests. An illustrative method involves categorizing items into "*head*" (*popular*) and "*tail*" (*less popular*) based on their popularity levels and tailoring separate datasets for generating recommendations for each category (Park and Tuzhilin 2008). Another approach by Jannach et al. involves creating user-item interactions where popular items are excluded to form pairs that emphasize lesser-known items, which are then used to train the recommender algorithms (Jannach et al. 2015). Furthermore, employing a probability distribution function that accounts for item popularity can enhance the visibility of underrepresented items in recommendation lists (Chen et al. 2018).

Efforts to address the issue of *popularity bias* in recommender systems have included the use of "*in-processing*" techniques, which adapt existing recommendation algorithms to simultaneously account for both relevance and popularity; hence, they are typically viewed as solutions specific to certain algorithms and not universally applicable. Key instances of such methods include: (i) a technique that assesses user preferences for less favored items and leverages this data to deprioritize *popular items* in the recommendation sequence (Kamishima et al. 2014); (ii) an enhanced version of the RankALS algorithm designed to maintain a balance between diversity within lists and overall accuracy (Abdollahpouri et al. 2017); (iii) a custom recommendation model that calculates the degree of familiarity among items based on popularity and selectively prunes highly preferred items to foster a more equitable common-neighbor similarity metric (Hou et al. 2018a); (iv) a *debiasing* strategy that reduces the correlation between the relevance of user-item pairs and the popularity of items, aiming for an equitable treatment of less mainstream, "*long-tail*" items (Boratto et al. 2021); (v) and an approach utilizing variational autoencoders that adjusts item scores based on historical popularity metrics to enhance the diversity of recommendation outcomes (Borges and Stefanidis 2021).

Conversely, post-processing techniques for addressing this challenge are adaptable to the outputs of any recommendation system, entailing the adjustment of rankings or the generation of a new recommendation list adhering to a particular constraint. A method suggested by Abdollahpouri et al. computes final scores by assigning weights to predicted ratings in inverse proportion to item popularity, subsequently utilizing these scores for the re-ranking of recommendations (Abdollahpouri et al. 2018). Likewise, Yalcin and Bilge introduce two distinct methods aimed at mitigating popularity bias within recommendations, specifically designed for user groups as opposed to individual users (Yalcin and Bilge 2021). These methods, named multiplicative and augmentative, both employ a re-ranking mechanism for items, yet they diverge in their methodology of penalizing items according to their popularity levels. The multiplicative method endeavors to diminish the overall popularity of items within recommendations by incorporating item weights directly into the calculation of ranking scores. Conversely, the augmentative method prioritizes prediction scores as the fundamental factor, incorporating item weights into ranking scores to ensure a satisfactory level of precision in the resultant recommendations. Another notable

strategy, the *xQuad* re-ranking method introduced by Abdollahpouri et al., aims to mediate the balance between accuracy in rankings and the representation of *long-tail* items (Abdollahpouri et al. 2019).

As outlined in the preceding sections, there's a burgeoning interest within the literature of Recommender Systems (RSs) in safeguarding *user privacy* and *mitigating the popularity bias* inherent in recommendation algorithms. Nevertheless, to knowledge, no studies have yet explored the impact of integrating privacy-protection mechanisms on the bias distribution within *Collaborative Filtering (CF)* recommenders. Furthermore, it is imperative to conduct additional investigations to determine if the effectiveness of *popularity-debiasing* techniques is maintained when traditional CF recommenders incorporate such *privacy-enhancing* strategies.

3.3. Popularity Debiasing Solutions

Recommender systems have evolved into ubiquitous tools within the daily lives of individuals, providing tailored recommendations for products, services, and content. These systems extensively employ data-driven algorithms to predict user preferences and deliver recommendations. However, there's growing concern over the vulnerability of these algorithms to *popularity bias*, where dominant items are disproportionately recommended over less popular or *niche* alternatives (Ahanger et al. 2022). Such bias can lead to the standardization of user preferences and diminish the *diversity* of recommendations. Recently, researchers have focused on developing strategies to mitigate the effects of *popularity bias* within recommendation algorithms (Yalcin 2022a; Chen et al. 2023). This section meticulously reviews the current literature on the issue of *popularity bias* and the solutions proposed to address it in recommender systems.

Initial inquiries into this domain primarily examine the structure and influence exerted by recommendation algorithms across various sectors, including online education (Boratto et al. 2019), film (Boratto et al. 2021; Borges and Stefanidis 2021), literature (Naghiaei et al. 2022), music (Jannach et al. 2016; Kowald et al. 2020), and travel (Sánchez 2019). A pioneering study conducts an exhaustive analysis of several leading recommendation techniques, with a focus on *collaborative filtering* algorithms, uncovering that their recommendations predominantly favor a small, popular subset of items (Jannach et al.

2015). This study further investigates how algorithmic formulation and parameter settings influence *popularity bias* and advocates for hyperparameter adjustments to foster diversity in recommendations. Additionally, another study assesses a collection of algorithms for their susceptibility to various biases, including the popularity within course categories, catalog coverage, and the popularity of courses in massive open online courses (Boratto et al. 2019). Findings indicate that these algorithms may greatly vary in their course recommendations and might display unwanted biases, thereby leading to potential educational consequences. Moreover, an associated study evaluates these representative algorithms effectively, demonstrating the effects of *popularity bias* on different stakeholders, such as individuals and item providers (Abdollahpouri 2020).

Numerous contemporary studies have ventured into analyzing the potential discriminatory effects engendered by *popularity bias*. These investigations primarily focus on understanding how the inclination of recommendation algorithms toward favoring popular items disproportionately affects diverse user groups or different entities within the ecosystem, such as consumers or content creators (Abdollahpouri 2020), or segments of users distinguished by age, ethnicity, or gender (Lesota et al. 2021). The seminal research on this issue reveals that individuals with a predilection for *less popular, niche* items encounter more pronounced adverse effects due to *popularity bias* in film recommendations (Abdollahpouri et al. 2019). This phenomenon of inequity has been similarly observed within the domains of book (Naghiaei et al. 2022) and music recommendations (Kowald et al. 2020). Furthermore, the issue has been analyzed concerning five vital attributes linked to user rating behaviors, indicating that users who are discerning, highly engaged, and exhibit unpredictable patterns receive disproportionately unfavorable recommendations, notwithstanding their substantial contributions to the platform (Yalcin and Bilge 2022). Moreover, a subsequent study investigates the unfairness through the lens of the big-five personality traits model, concluding that individuals with lower levels of extroversion or a reluctance towards new experiences face more biased recommendations regarding item popularity (Yalcin and Bilge 2023).

In the realm of literature, a multitude of pragmatic approaches has been formulated to tackle the challenge of bias dissemination within recommender systems. These approaches are typically segregated into three principal categories: *post-processing*, *in-*

processing, and *pre-processing*, contingent upon their method of integration during the phase of recommendation production.

Post-processing debiasing techniques aim to adjust or recreate ranked lists from conventional recommendation systems by demoting popular items and promoting *less popular* ones. These techniques are highly versatile, capable of being applied to the output from any recommendation system, thus offering broader applicability than alternative methods. For instance, one approach recommends employing predefined user-specific weights to reduce the impact of *popularity bias* (Jannach et al. 2015), which also helps in maintaining acceptable levels of accuracy. Another strategy involves generating synthetic ranking scores by weighting predictions inversely to the item's popularity (Abdollahpouri et al. 2018), leading to the arrangement of items based on these scores to form the final recommendations. In an effort to counteract the negative implications of *popularity bias* in group recommendation scenarios, have previously introduced two improved re-ranking strategies, Augmentative and Multiplicative, each designed to penalize popular items (Yalcin and Bilge 2021). The Augmentative method focuses on enhancing ranking accuracy, while the Multiplicative method is aimed at including more unpopular items in the recommendations. Moreover, a novel debiasing approach has been proposed to ensure fairness in item exposure within recommendation systems by implementing constraints based on item stock volumes (Dong et al. 2021). This method limits the recommendation frequency of an item relative to its historical interaction, demonstrating through empirical evidence its superiority over traditional recommenders by effectively mitigating the *Matt-hew Effect* that skews item popularity. This heuristic method leverages normalized scores and a *Minimum Cost Maximum Flow* approach, considering both the recommendation length for users and item stock levels. Consequently, it effectively tackles both the user-item matching problem and the issue of *popularity bias* in the final recommendations.

Additionally, a significant study introduces a highly effective technique for promoting *long-tail* items, derived from the *xQuad* algorithm, aimed at improving recommendation diversity (Abdollahpouri et al. 2019). Another contemporary method designed to reduce popularity bias, known as *Discrepancy Minimization*, seeks to enhance overall diversity in recommendations through a *minimum-cost network flow* strategy, targeting the exploration of recommendation sub-graphs that maximize diversity (Antikacioglu and Ravi 2017). In

contrast, *the FA*IR* method, an alternative approach to mitigating popularity bias, focuses on establishing an equilibrium between popular and less popular items by creating separate queues for these categories and then merging them (Zehlke et al. 2017). Lastly, according to current knowledge within the academic literature, Calibrated Popularity is the unique method that considers users' inherent preferences towards item popularity in the debiasing process (Abdollahpouri et al. 2021). This technique initially identifies users' true interests in three categories of items - *head*, *middle*, and *tail* - based on the quantity of ratings they have received. This data is then utilized in calculating the *Jensen-Shannon divergence* for generating new ranked recommendation lists. However, a notable limitation of this method is its consideration of all ratings in users' profiles without assessing their actual values in determining users' interest levels in popularity, overlooking the fact that users might rate popular items negatively and that a rating does not necessarily indicate a user's enjoyment of an item.

In the realm of recommendation systems, *in-processing* methods for reducing popularity bias aim to adjust the core functionality of a recommendation algorithm to neutralize its inherent tendency towards favoring popular items. These techniques are identified as non-universal solutions due to their specificity to certain recommendation systems and their inapplicability across different platforms. One such method evaluates the probability of users disfavoring items and leverages this insight to impose penalties on popular items during the recommendation prediction process (Kamishima et al. 2014). Moreover, a recent innovation has led to the development of an optimized version of the *RankALS* recommender, which strives for enhanced diversity within recommendation lists and improved accuracy in rankings (Abdollahpouri et al. 2017). Additionally, another study proposes a framework that predicts the common neighbors between items based on their initial popularity levels, then filters out the most popular ones to achieve a balanced similarity index among co-neighbors (Hou et al. 2018a). Another strategy targets the reduction of popularity bias by decreasing the correlation between the relevance scores of user-item pairs and the popularity of items, thereby ensuring a fairer representation of less popular items (Boratto et al. 2021). A novel approach grounded in *Variational Autoencoders* seeks to decrease bias and increase diversity in recommendations by penalizing items based on their past popularity (Borges and Stefanidis 2021). Lastly, a groundbreaking approach

ach named *Popularity-bias Deconfounding and Adjusting (PDA)* has been introduced to eliminate the confounding effects of popularity bias during the training phase of models and to fine-tune recommendation scores with a controlled level of popularity bias during inference (Zhang et al. 2021).

The manifestation of bias within recommendation systems is frequently linked to the disproportionate distribution of user ratings. Consequently, *pre-processing* strategies are designed to rectify this imbalance by altering the dataset on which recommender systems are trained. One such strategy involves segregating items into "*head*" or "*tail*" categories based on the volume of ratings they receive. Predictions for head items are generated using their complete rating data, whereas only ratings from the same category are employed for tail items to prevent the overshadowing of these items by more popular ones (Park and Tuzhilin 2008). Another tactic in this vein is the creation of user-item pairs from which popular items are intentionally omitted. These pairs are then used as the training set for recommendation algorithms, rather than the full user-item rating matrix, to ensure a more equitable representation of items (Jannach et al. 2015). Furthermore, an additional method employs a probability distribution function that accounts for an item's popularity, enhancing the chances for less popular, or tail, items to be recommended by incorporating them into the ranking more frequently (Chen et al. 2018).

A novel study tackles the issues of bias towards popularity and unfairness within recommendation systems by focusing on user interactions with items of high popularity. The introduced algorithm mitigates these biases and enhances the fairness of recommendations, alongside a marginal improvement in the accuracy of recommendations, by filtering out what are termed as "*unreliable interactions*" (Ma and Dong 2021). Another research introduces a model designed to diminish the bias in recommendations and ensure their utility by examining the impact of herd behavior on user ratings (Su et al. 2023). Furthermore, a recent publication unveils an algorithm aimed at curtailing the bias towards popular items in recommender systems (Gangwar and Jain 2021), which typically results in an undue emphasis on popular items at the expense of lesser-known ones. This algorithm recalibrates the significance attributed to less popular items, thereby enhancing their visibility. Comprehensive evaluations have shown this technique to be effective in reducing the bias towards popularity without compromising the accuracy of the recom-

mendations. Lastly, the introduction of the TASTE model, which utilizes textual representations to align items with users, marks an advancement in improving the precision of recommendations (Liu et al. 2023). *TASTE* significantly diminishes popularity bias and tackles the challenge of the cold start by employing comprehensive text modeling and leveraging pre-trained language models, leading to recommendations that are both more pertinent and fitting.

As discerned from the literature review provided, extensive efforts have been directed towards *mitigating popularity bias*. For example, post-processing techniques, which are applied after a recommender system has generated its predictions, might lead to a reduction in accuracy and an increment in system complexity. Moreover, such methods might only offer superficial solutions as they do not tackle the inherent bias within the model but merely modify its outputs. On the other hand, *in-processing methods*, implemented during the model's training phase, could enhance the model's complexity and prolong the training time, potentially altering the data structure. In contrast, *pre-processing strategies* can circumvent these drawbacks by addressing biases in the dataset prior to the commencement of the training process for the recommender system. These approaches act on the data before the training phase, thus diminishing bias without exacerbating model complexity and preserving the accuracy of the system. Therefore, *pre-processing techniques* may present a more viable and efficient solution for countering popularity bias in recommendation systems as compared to *post-processing* and *in-processing methods*.

Moreover, current pre-processing approaches to tackling the issue of popularity bias are frequently assessed using a narrow set of metrics, predominantly concentrating on either accuracy or the extent of popularity bias. Nonetheless, there exists a multitude of dimensions beyond mere accuracy that affect the efficacy of recommendation systems. These include attributes like *novelty*, *diversity*, *entropy*, and coverage of the *long-tail items*, all of which may be influenced by popularity bias. Hence, there is a pressing need for more comprehensive strategies that can address these factors simultaneously. Introduced solution, *EquiRate*, fulfills this requirement by not just equalizing the distribution of ratings across different items through the introduction of ratings that reflect genuine patterns and preferences found in user data but also by producing recommendation lists that simultaneously satisfy a range of criteria, including accuracy and beyond accuracy.

4. MATERIALS AND METHODS

At this stage of the research, the effects of privacy prevention methods in the literature on the popularity bias problem were discussed. In addition, the answers to the two research questions mentioned in the first chapter are sought in this section.

4.1. Constructing User Personas Based on Varying Privacy Protection Requirements

Gathering precise data from users is essential for the effectiveness of *Collaborative Filtering (CF)* techniques in generating high-quality recommendations. Nonetheless, concerns over privacy often deter individuals from disclosing their genuine preferences regarding items or products. This reluctance complicates the accurate modeling of user preferences, consequently resulting in substandard recommendations. Hence, the challenge of delivering precise recommendations while preserving user privacy has emerged as a pivotal issue in the realm of Recommender Systems (RSs) in recent times (Ozturk and Polat 2015; Pramod 2022).

Privacy-Preserving Collaborative Filtering (PPCF) strategies are primarily designed to obscure (i) the inventory of items that users have actually rated and (ii) the authentic values of those ratings. The concealment of the item list can shield users from malevolent activities, such as targeted unsolicited marketing, whereas hiding the true rating values can defend against potential price discrimination or profiling efforts. Among the foremost strategies for safeguarding user privacy, *Random Perturbation Techniques (RPTs)* are deployed to systematically insert fictitious votes into vacant slots within user profiles and to introduce random alterations to each real rating present (Polat and Du 2003; Yargic and Bilge 2019). Thus, these methodologies effectively mitigate the risk of exposing sensitive user data, offering a commendable balance between maintaining the utility of preference data and ensuring privacy without significantly compromising on accuracy.

In more specific terms, these methodologies propose masking a rating entry v by substituting it with $v + r$, where r represents a randomly selected value originating from either a *Gaussian* or *uniform distribution*. The spectrum of these random values is modulated by a predetermined σ_u parameter, where $\sigma_u = [0, \sigma_{max})$ and σ_{max} is set by the central server to manage the distortion degree of the genuine user preferences. Assuming α is

a fixed value equating to $\sqrt{3}\sigma_u$, the random values are sourced from either $\mathcal{U}(-\alpha, +\alpha)$ or $\mathcal{N}(0, \sigma_u^2)$ distributions. Furthermore, RPTs employ these random values as fictitious ratings to fill arbitrarily chosen vacant slots within user profiles. The quantity of these slots to be filled is dictated by a chosen β_u parameter, where $\beta_u = [0, \beta_{max})$ and β_{max} is established by the central server in relation to the density proportion of the complete user-item rating matrix.

Algorithm 1 Randomized perturbation procedure for data concealing

Input: User rating vector ($u_{1 \times n}$), σ_{max} , and β_{max}

Calculating z-scores ($\rightarrow Z$):

1: **for all items** in u ($j \leftarrow 1$ to n) **do**

2: $z_j = [u_j - \text{MEAN}(u)] / \text{STD}(u)$

3: **end for**

Adjusting privacy parameters:

4: $\sigma_u \leftarrow \text{RAND}(0, \sigma_{max})$, $\beta_u \leftarrow \text{RAND}(0, \beta_{max})$, and $\alpha \leftarrow \sqrt{3}\sigma_u$

5: $e \leftarrow \#$ of empty cells, $g \leftarrow \#$ of genuine ratings

6: $F \leftarrow e \times \beta_u \%$

Selecting dispersion:

7: $dist \leftarrow \text{RAND}(\text{Gaussian}, \text{uniform})$

Creating random values:

8: **if** $dist$ is Gaussian:

9: $R \leftarrow \mathcal{N}(g + F; \mu = 0, \sigma_u)$

10: **else:**

11: $R \leftarrow \mathcal{U}(g + F; \alpha)$

12: **end if**

Disguising z-scores ($\rightarrow Z'$):

13: **for all items** in u ($j \leftarrow 1$ to n) **do**

14: $z'_j = (z_j + R_j)$

15: **end for**

return Z'

Figure 4.5. Randomized perturbation procedure for data concealing

In this context, each user customizes their σ_u and β_u settings according to the preferred level of privacy and then obscures their profiles following the procedure outlined in figure 4.5. Subsequently, they distribute their camouflaged profiles in place of the genuine ones to the service provider. As a result, a camouflaged preference matrix, denoted as $R'_{m \times n}$, is formed, where m and n represent the total number of users and items, respec-

tively. This matrix serves as the foundation for generating recommendations through any suitable CF recommender.

To thoroughly evaluate the impact of popularity bias on users with differing privacy concerns, categorize users into four groups based on their preferred level of privacy protection through the implementation of figure 4.5. The first category, named *Utterly Unconcerned (UU)*, comprises individuals who adhere to conventional non-private CF methods, displaying total indifference towards their privacy and expecting the highest accuracy in recommendations. The second category, *Barely Bothered (BB)*, consists of users who opt for minimal privacy protection and remain untroubled by minor privacy breaches. The third group, *Pretty Pragmatists (PP)*, includes those who seek a practical compromise between privacy protection and recommendation accuracy, wishing for satisfactory recommendations without compromising their privacy. Lastly, the *Fiery Fundamentalists (FF)* category is reserved for users extremely vigilant about their privacy, prioritizing it over the accuracy of the recommendations they receive.

In ensuring user privacy through the application of *RPTs*, the chosen values for σ_{max} and β_{max} play a pivotal role in customizing their preferred level of privacy. Notably, higher (or lower) values of σ_{max} and β_{max} correspond to a more (or less) enhanced level of privacy protection. Hence, delineate these four user personas with specific attention to the parameter adjustments outlined in Table 4.4 while implementing *RPTs*. This allows us to scrutinize each persona individually, providing a clearer understanding of the negative repercussions associated with each level of privacy-centric user personas.

Table 4.4. Parameters for the devised privacy-centric user profiles

User Persona	σ_{max}	β_{max}	Dispersion	Privacy-level
<i>Utterly Unconcerned (UU)</i>	0	0	N/A	No privacy
<i>Barely Bothered (BB)</i>	2	5	Uniform/Gaussian	Low-level
<i>Pretty Pragmatists (PP)</i>	3	10	Uniform/Gaussian	Mid-level
<i>Fiery Fundamentalists (FF)</i>	4	25	Uniform/Gaussian	High-level

While the privacy levels provided are intuitively evident in Table 4.4, further introduce

a method to quantify the specific degree of privacy achieved through the β_{max} and σ_{max} parameters. To accomplish this, conduct distinct sets of analyses aimed at quantifying the privacy levels facilitated by these parameters.

As delineated by figure 4.5, each anonymized user profile amalgamates authentic and spurious ratings, occupying $\beta_{max}\%$ of the vacant cells. Consequently, any breach of privacy by the server necessitates the preliminary identification of the actual items between some level of uncertainty. This uncertainty can be pragmatically evaluated by calculating the likelihood that any given rating is authentic. Let $a = \{a_1, a_2, \dots, p_a\}$ represent the array of genuine ratings and $f = \{f_1, f_2, \dots, f_{e \times \beta_{max}}\}$ denote the assortment of inserted fictitious ratings. The probability that a randomly chosen rating, r , is authentic, denoted as $\Pr(r \in a)$, can thus be approximated as $\frac{|a|}{|a|+|f|}$. It's noted that a lower probability signifies a higher degree of privacy safeguarding, and the inverse implies a reduction in privacy protection.

Histograms showcasing the estimated probabilities for three user personas across three datasets (namely, *MovieLens-1M* (ML), *Douban Book* (DB), and *Yelp*) are depicted in Fig. 4.6, with the specifics of these datasets elaborated in the subsequent section. As anticipated, *FF* users predominantly occupy the bins with the lowest probabilities in the histograms, whereas *PP* and *BB* users exhibit a distribution towards bins with higher probabilities. This distribution pattern underscores a gradational diminution in confidentiality levels from *FF* to *BB*. It is noteworthy that the impact of β_{max} on the privacy level intensifies as the ratio of the number of ratings per user to the total number of items in the dataset diminishes, a trend that becomes particularly evident in larger datasets like *Yelp* and *DB*.

In the realm of rating distribution, assessing the level of privacy achieved by incorporating random noise to separate real ratings from fabricated ones is of paramount importance. While central servers are capable of distinguishing between these, it is imperative to extract the true values from their perturbed z -score representations. To evaluate the privacy level of a variable perturbed with additive noise, Agrawal and Aggarwal introduce a measure grounded in differential entropy (Agrawal and Aggarwal 2001). This measure has been adopted in the *PPCF* framework by (Polat and Du 2005b; Bilge and Polat 2013). Within this framework, the initial user vector is denoted as P , and the perturbing random

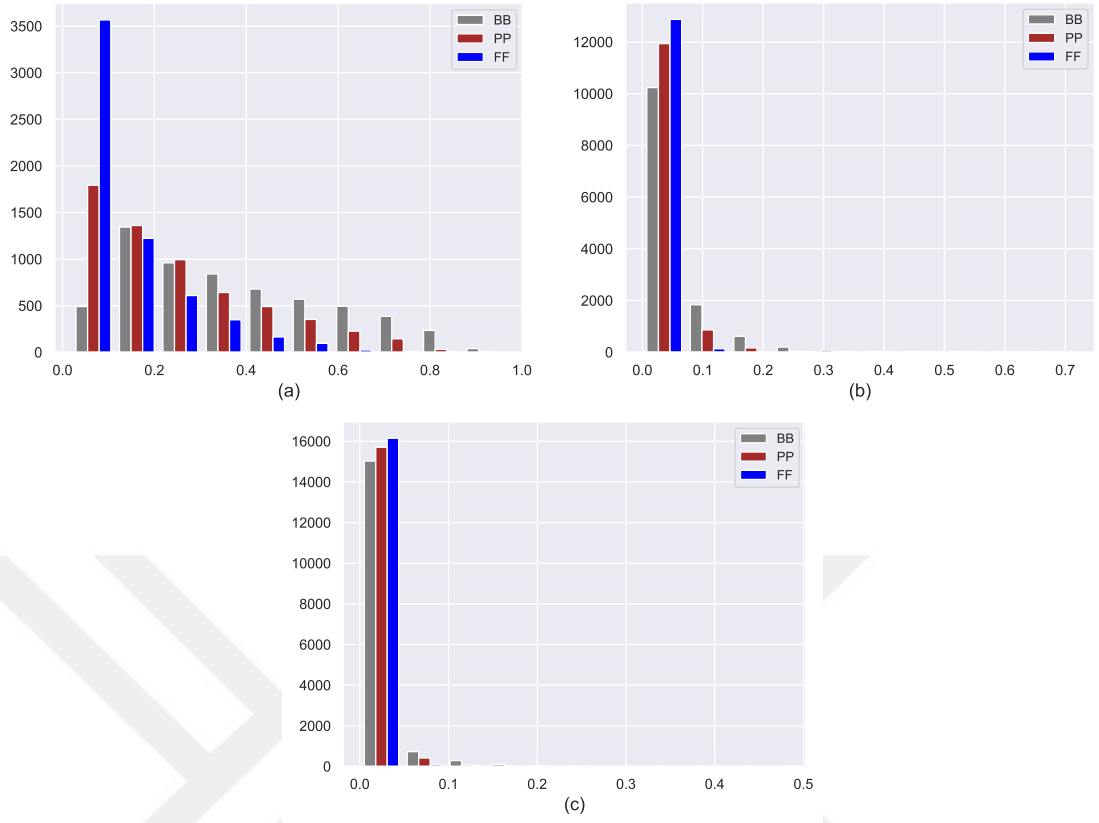


Figure 4.6. Histograms depicting the estimated probabilities for three personas within the (a) ML, (b) DB, and (c) Yelp datasets.

noise as R , resulting in $U = P + R$. The *average conditional privacy* of P is delineated as $\Pi(P|U) = 2^{H(P|U)}$, where $2^{H(P|U)}$ signifies the *conditional differential entropy* of P conditional on U . Given that P and R are independent random variables, the privacy metric for P upon the revelation of U is calculated as $\Pi(P|U) = \Pi(P) \times (1 - Pr(P|U))$, with $Pr(P|U) = 1 - 2^{H(U|P) - H(U)}$. Assuming P follows a normal distribution, the privacy levels, $\Pi(P|U)$, across different perturbation intensities are depicted in Table 4.5. The findings elucidate that privacy levels ascend in tandem with the augmentation of profile perturbation, as anticipated.

4.2. Experimental Setup

In this section, delineate the experimental framework, encompassing the datasets employed, the algorithms for recommendations, and the protocols for evaluation. Additionally, outline the methodology for the experiments and the approaches considered for address-

Table 4.5. Privacy levels ($\Pi(P|U)$) for different personas

Dataset	User Personas		
	<i>BB</i>	<i>PP</i>	<i>FF</i>
ML	2.644	2.883	3.021
DB	1.864	1.985	2.051
Yelp	1.464	1.540	1.579

sing the issue of popularity bias.

4.2.1. Evaluation Metrics

In this study, aim to evaluate the influence of *perturbation* and *obfuscation* techniques on the prevalence of popularity bias from various angles, hence considering nine crucial evaluation metrics throughout the experiments. Among these, the *Average Percentage of Recommended Items (APRI)* metric, proposed by Yalcin, exclusively quantifies the popularity level of items recommended (Yalcin 2022b). The rest of the metrics are categorized into two principal groups: *accuracy* and *beyond-accuracy* metrics. Specifically, for accuracy assessment, utilize well-established metrics such as *nDCG*, *Precision*, *Recall*, and *F1-score*. For evaluating the quality of recommendations beyond accuracy, adopt *Average Percentage of Long-tail Items (APLT)*, *Novelty*, *Long-tail Coverage (LTC)*, and *Entropy*. While the chosen beyond-accuracy metrics are designed to evaluate various recommendation aspects like *popularity*, *diversity*, *coverage*, *novelty*, and *inequality*, they all ultimately relate to the well-acknowledged concept of fairness in the recommendation system (RS) literature from an item perspective (Abdollahpouri and Burke 2019). Below, elaborate on these metrics in detail, where $i_1, i_2, \dots, i_N$ represents the items in the top- N list for a user u , denoted by N_u , as generated by a recommendation algorithm.

4.2.2. Experimentation Methodology

In analysis, examine four distinct scenarios as outlined in Table 4.4, characterized by users' varying levels of privacy protection preferences. It is important to note that the experiments described subsequently are conducted independently for each of these

scenarios.

For each defined user profile scenario, top-10 recommendations are generated for users utilizing one of the algorithms previously described. This is achieved by implementing the renowned *leave-one-out cross-validation* technique (Kearns and Ron 1997). In this method, a specific user is designated as the test case while the rest serve as the training data. The algorithm in question is executed on the training dataset to make predictions for all possible items for the test user. This procedure is iteratively applied to every user within the dataset. Subsequently, the top-10 items are chosen for each user to form the recommendation list, sorted based on their prediction scores in a descending order.

To conduct a thorough evaluation on the impact of recommendations on users with different privacy concerns, calculate the *Precision*, *Recall*, *F1-score*, *nDCG*, *APRI*, and *Novelty* metrics for the top-10 recommendations of each user and then compute their average values to derive the overall results. Conversely, for the assessment of the *LTC* and *Entropy* metrics of the algorithms, aggregate the top-10 recommendations of all users, as these measurements are not focused on individual users but rather evaluate the entire recommendation set collectively. It's important to note that in the subsequent section where, present the experimental results, take into account the average outcomes of 100 iterations conducted for each user profile scenario to mitigate potential variances introduced by the randomization in the *perturbation* and *obfuscation* stages.

Each of these algorithms offers unique advantages and is suited to different types of data and recommendation scenarios. The choice of algorithm depends on the specific requirements and characteristics of the application, including the nature of the feedback (*explicit* or *implicit*), the scale of the dataset, and the desired balance between accuracy and complexity.

Enhancing the experiments with various parameters is crucial for a deeper analysis of potential fairness issues among users with different privacy preferences. Utilizing diverse algorithms in the phase of generating recommendations is the most effective strategy for this aim. This methodology involves creating recommendation lists through various algorithms subsequent to data *perturbation* and *obfuscation* techniques, to assess the impact of privacy protection measures on the propensity of each algorithm towards popularity

bias.

Consequently, ten diverse algorithms have been identified for recommendation purposes, encompassing two non-personalized and eight cutting-edge methods across four distinct CF families (namely, based on *matrix factorization*, *probabilistic models*, *clustering*, and *neural networks*), as elaborated in Table 2.2. Two non-personalized techniques, specifically *ItemAverage* (*IA*) and *MostPopular* (*MP*), were chosen for their prevalent use in scholarly works for generating basic recommendations (Boratto et al. 2019; Yalcin and Bilge 2022). In particular, *IA* advises items based on the highest mean ratings, while *MP* ranks items by the volume of their ratings in a descending order, prioritizing the most frequently rated items for recommendations. Thus, inherently, *MP* yields identical recommendation lists for every user, signifying the epitome of popularity bias that could be perpetuated. Conversely, the remaining eight CF algorithms, representing the top performers from their respective families, are implemented through a Python-based library known as Cornac¹⁰ (Salah et al. 2020). Moreover, for the sake of reproducibility, the algorithms' hyperparameters were maintained at their default settings. It's important to note that the primary aim of this study is to examine the propagation of disparities among user personas based on different privacy levels by recommendation algorithms, hence detailed descriptions of these algorithms' workings are omitted for clarity.

4.2.3. The Benchmark Treatment Methods for Popularity Bias Problem

One of the principal aims of this research is to investigate the variability in the efficacy of conventional popularity bias mitigation strategies when they are implemented on anonymized user profiles. For this purpose, assess three established *post-processing* techniques dedicated to addressing the *popularity bias* in recommendation systems, focusing on various attributes of recommendation quality. These techniques endeavor to adjust the ranking of recommended items by imposing penalties on popular items. They incorporate a tuning parameter (denoted as λ) to balance the trade-off between prediction *accuracy* and a significant *beyond-accuracy* metric such as diversity or coverage.

The two popularity-debiasing techniques implemented in this study are adaptations of the *Enhanced Re-ranking Procedure* outlined by (Yalcin and Bilge 2021): namely,

¹⁰<https://cornac.preferred.ai/>

Augmentative (Aug) and *Multiplicative (Mul)*. These methods reorder the items within recommendation lists according to computed synthetic scores, which are derived by inversely adjusting their predictive scores and popularity metrics. However, the approaches distinguish themselves through their score calculation strategies; specifically, *Mul* aggressively discounts the scores of popular items by inversely multiplying their predictive scores with their popularity metrics, whereas *Aug* leverages the predictive scores as the primary factor, with item popularities contributing as secondary modifiers. As a result, the *Mul* method can significantly enhance the beyond-accuracy aspects of recommendations, potentially at the cost of reduced accuracy. Conversely, the *Aug* method prioritizes maintaining predictive accuracy while addressing the influence of popularity bias to the greatest extent feasible.

The third algorithm employed in this research is the notable *xQuad (xQd)* algorithm introduced by Abdollahpouri et al., dedicated to generating user-specific final recommendation lists through a dynamic selection process of items categorized into: (i) *head*, representing popular items, and (ii) *tail*, denoting the less popular or *niche* items (Abdollahpouri et al. 2019). Initially, the algorithm selects a pool of candidate items solely based on their prediction scores yielded by a preceding recommendation algorithm. Subsequently, it assembles the ultimate recommendation lists by dynamically adjusting the ranking scores of these candidate items. These adjustments are made in each iteration by modulating their prediction scores, taking into account both the user’s inherent inclination towards the item’s category (i.e., *head* or *tail*) and the representation ratio of each category within the current recommendation list.

4.3. Analyzing Robustness of Privacy-Preserving Collaborative Recommenders Against Popularity Bias Issue

The initial series of experiments conducted focuses on assessing the extent to which the popularity bias within recommendation systems impacts the quality of recommendations for the baseline scenario, *UU*, along with three user privacy protection scenarios, namely *BB*, *PP*, and *FF*. Essentially, evaluate the differential impact of algorithmic bias propagation on users with varying levels of privacy concerns. For this purpose, orchestrate a variety of experiments leveraging ten distinct recommendation algorithms (*MP*, *IA*,

MMMF, *WMF*, *HPF*, *IBPR*, *WBPR*, *SKM*, *NEUMF*, and *VAECF*) and three datasets (ML, DB, and Yelp), as previously discussed, and calculate the outcomes for nine metrics across four privacy-centric scenarios. Consequently, Figures 4.7 - 4.13 display the results for beyond-accuracy metrics, namely *APRI*, *APLT*, *Entropy*, *Novelty*, and accuracy metrics, namely *Precision*, *Recall*, *F1-score*, and *nDCG*, for the three datasets, respectively.

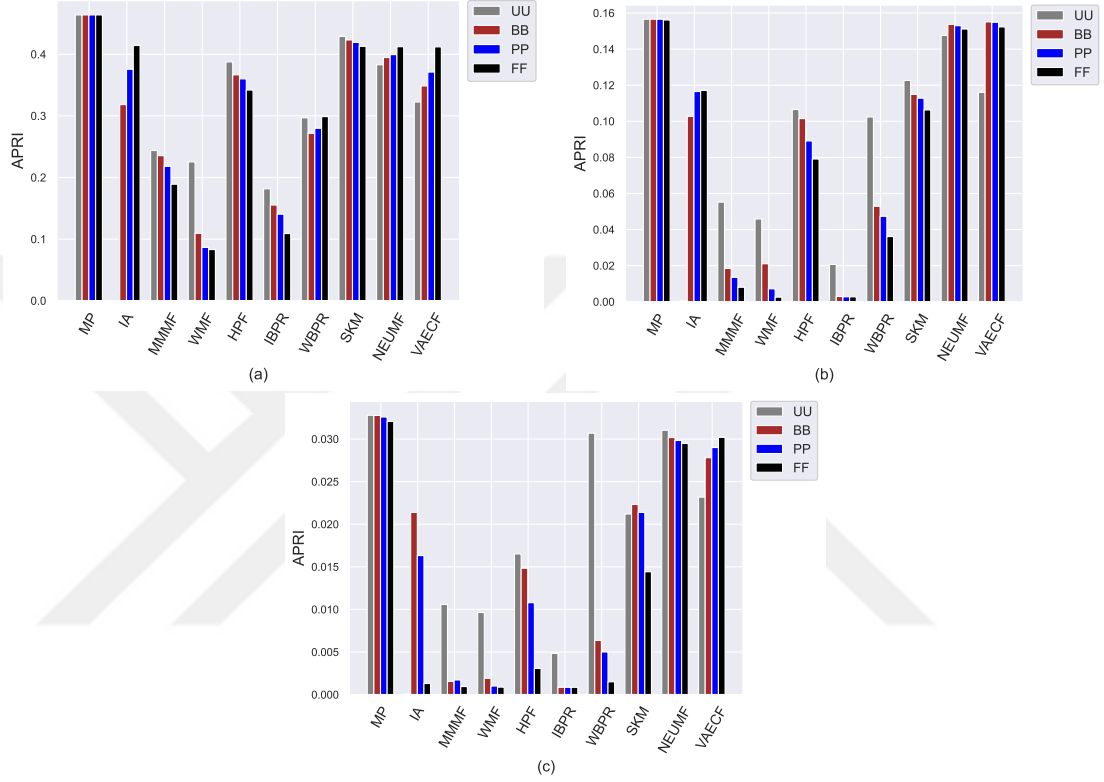


Figure 4.7. The *APRI* outcomes derived from the algorithms applied across user profiles categorized by four distinct levels of privacy within the (a) ML, (b) DB, and (c) Yelp datasets.

As depicted on the y-axis of Fig. 4.7, irrespective of the algorithm used and the privacy level-specific user profiles, the highest *APRI* scores are typically recorded for the ML dataset, followed by the DB and Yelp datasets in sequence. This pattern is primarily attributed to the fact that the ML dataset features a higher average number of ratings per item compared to the DB and Yelp datasets, as detailed in Table 2.3. Such a scenario naturally leads to an elevated popularity ratio for items and consequently higher *APRI* values for the ML dataset. The *MP* algorithm is chosen to illustrate the maximum propagation of popularity bias in recommendations. Hence, as anticipated, the most sig-

nificant popularity bias in recommendations is achieved with the *MP* algorithm across all configurations. Additionally, an unforeseen discovery is that the neural networks-inspired algorithms, namely *NEUMF* and *VAECF*, achieve almost identical *APRI* outcomes as the *MP* algorithm, especially in the DB and Yelp datasets. Moreover, the *SKM* algorithm can be regarded as one of the least effective algorithms in terms of *APRI* outcomes for the ML dataset. Therefore, it can be concluded that despite being advanced personalized CF algorithms, they unfortunately exhibit extreme bias towards popular items. On the contrary, the lowest *APRI* values are generally observed with the *IBPR* or *WMF* among the algorithms used, particularly in the sparser and larger DB and Yelp datasets, indicating their high resilience to the popularity bias issue.

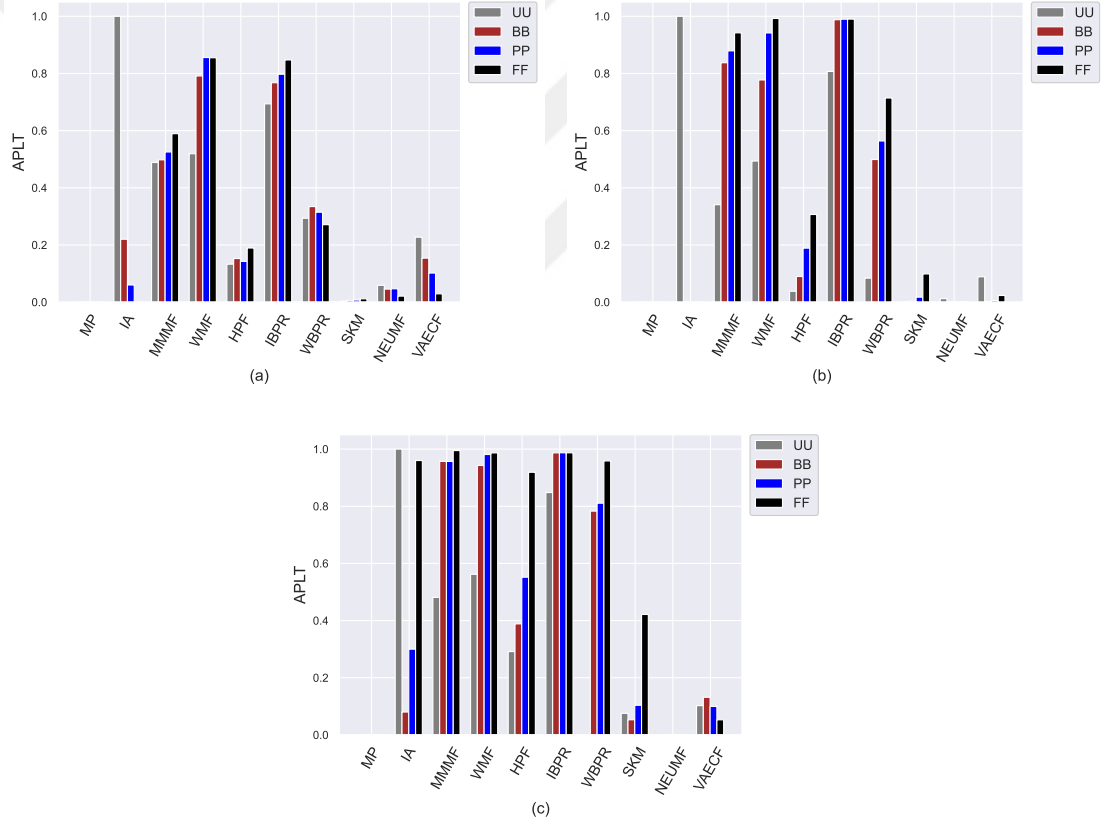


Figure 4.8. The *APLT* outcomes derived from the algorithms applied across user profiles categorized by four distinct levels of privacy within the (a) ML, (b) DB, and (c) Yelp datasets.

The most significant observation derived from Fig. 4.7 is that the lowest *APRI* scores are typically seen in the *FF* scenario, followed by *PP*, *BB*, and *UU*, indicating a decrease

in the average popularity of recommended items as the privacy level of users escalates. This trend holds true across almost all instances, with few exceptions. The variance in *APRI* scores across different privacy levels can be substantial in certain cases. For instance, within the Yelp dataset, the *APRI* score for *UU* is nearly five times that of *BB* when using the *WBPR* algorithm, or *PP* is almost four times higher than *FF* with the *HPF* algorithm. Hence, it can be inferred that privacy-enhancing measures in CF algorithms do more than just address users' privacy concerns; they also enhance the quality of recommendations by diminishing the algorithms' bias towards popular items. This effect is more pronounced for the DB and Yelp datasets compared to the ML dataset, signifying an important insight given that data collections in real-world applications are generally sparser and larger.

Contrary to the values of *APRI*, the results for *APLT* do not differ significantly across the datasets, as illustrated in Fig. 4.8. This consistency is attributed to *APLT* being a user-centric metric that gauges the average proportion of *long-tail* items in the top-10 recommendations. Given the earlier *APRI* findings, which identified *MP*, *SKM*, *NEUMF*, and *VAECF* as particularly susceptible to *popularity bias*, it is unsurprising that these algorithms struggle to highlight *long-tail (niche)* items across varying levels of user privacy in their recommendations. Conversely, *WMF* and *IBPR* stand out for their ability to diversify recommendations, evidenced by their superior *APLT* scores. This advantage is especially pronounced in the sparser and larger datasets of DB and Yelp, potentially due to the larger tail item set size compared to ML, which increases the likelihood of tail items being recommended to users in these datasets.

Another pivotal observation from Fig. 4.8 highlights that, with a few exceptions primarily related to CF approaches influenced by neural networks, the highest *APLT* scores are typically secured by *FF* users, succeeded by *PP*, *BB*, and lastly *UU*. This finding emphasizes the direct correlation between the escalation of user privacy and the enhancement in the diversity of the recommendations they receive. Furthermore, the disparity in *APLT* scores is more pronounced for the Yelp and DB datasets. The primary cause of this effect is attributed to the augmentation of user privacy through the integration of fictitious ratings into their profiles, leading to substantial alterations in rating distributions. Such adjustments render the algorithms more resilient to popularity bias, as evidenced in the

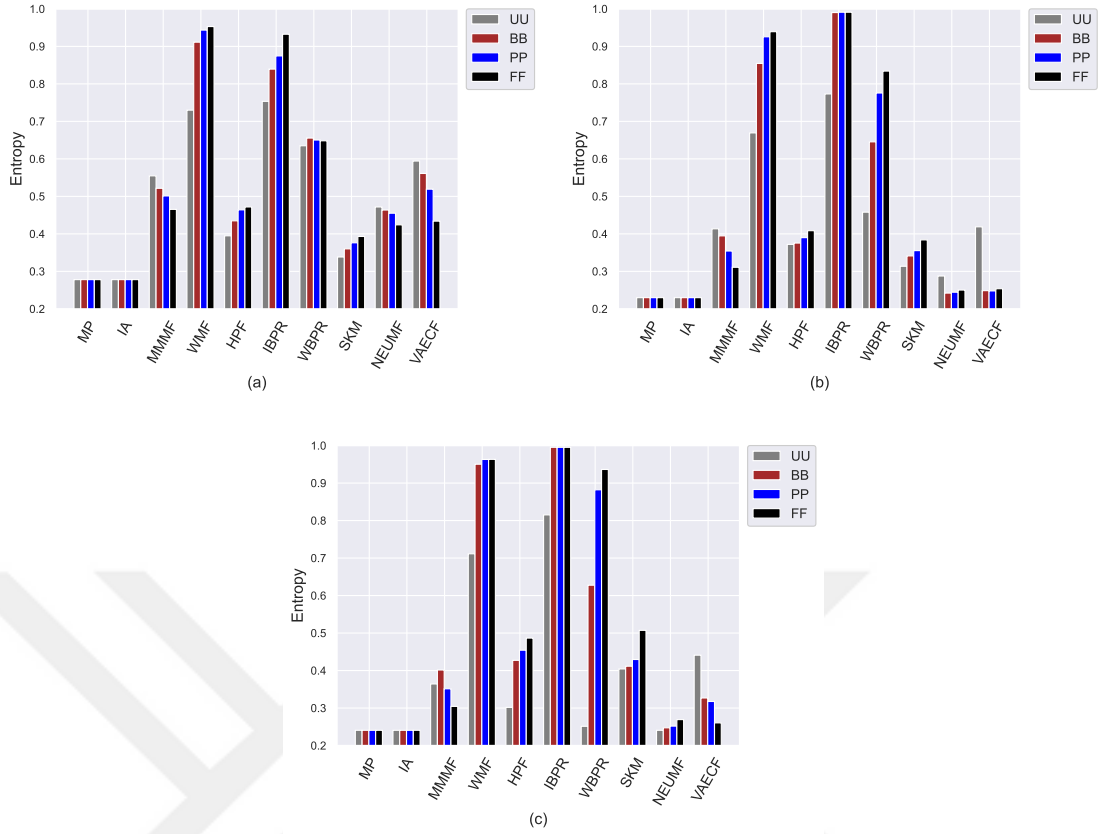


Figure 4.9. The *Entropy* outcomes derived from the algorithms applied across user profiles categorized by four distinct levels of privacy within the (a) ML, (b) DB, and (c) Yelp datasets.

prior discussion on *APRI* outcomes. Consequently, this mechanism naturally increases the likelihood of incorporating *long-tail* items into the ultimate recommendations.

In concordance with the *APLT* values obtained, the scores for *Entropy*, *Novelty*, and *LTC* assigned to the algorithms also span the same scale, that is, the $[0, 1]$ interval, across all data sets irrespective of the user profiles defined by their privacy level, as illustrated by Figs. [4.9] - [4.11], respectively. It is further noted from Fig. 4.9 that the minimal information entropy outcomes are predominantly associated with the *SKM*, *NEUMF*, and *VAECF* algorithms. Consequently, it is inferred that algorithms inspired by neural networks exhibit a significant deficiency in evenly distributing candidate items across the catalogue in their recommendations.

In comparison to other algorithms, the limitations of the *SKM*, *NEUMF*, and *VAECF* algorithms become evident in their capacity to provide recommendations with high no-

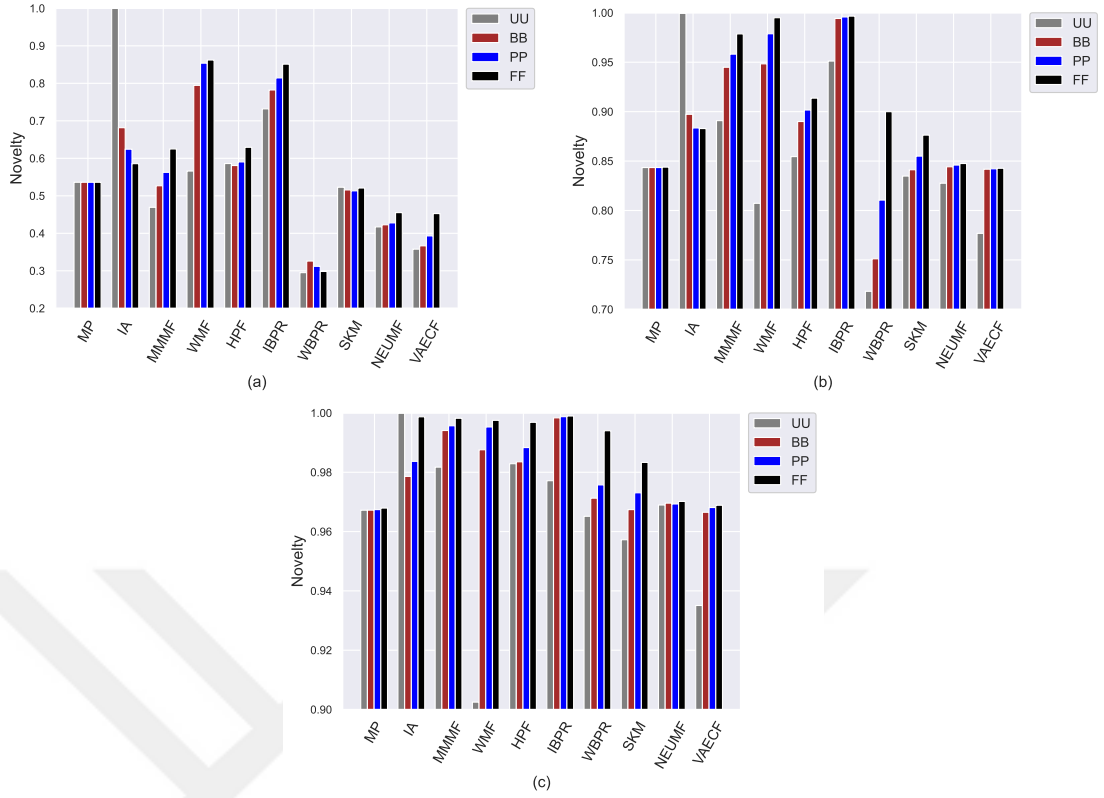


Figure 4.10. The *Novelty* outcomes derived from the algorithms applied across user profiles categorized by four distinct levels of privacy within the (a) ML, (b) DB, and (c) Yelp datasets.

velty and catalogue coverage, as demonstrated in Figs. 4.10 and 4.11. These algorithms typically exhibit the lowest *Novelty* and *LTC* scores across different privacy level-based user personas considered. Additionally, given that their *LTC* scores are generally near zero, it is evident that these algorithms, including *HPF*, are markedly inefficient for item-/service providers who aim to promote every product at least once through recommendations that span the entire catalogue. Furthermore, it is observed that the *Novelty* scores for all algorithms, even those based on *neural networks*, are lower for the ML dataset compared to the DB and Yelp datasets. This discrepancy is attributed to the average number of ratings per user being approximately 166 for ML, in contrast to 61 and 12 for the DB and Yelp datasets, respectively, as indicated by Table 2.3. Such variances in the datasets enable recommendation algorithms to more effectively identify items that are unknown or unexperienced by users in the DB and Yelp datasets, thus resulting in higher *Novelty* scores for these datasets, which are characterized by their sparser and larger nature.

Conversely, the *IBPR* and *WMF* algorithms, particularly the former, consistently achieve the highest *Entropy*, *Novelty*, and *LTC* metrics across all evaluated datasets and user personas categorized by privacy level, as illustrated in Figs. [4.9 - 4.11]. In essence, the recommendations produced by these two algorithms excel in equitable coverage and representation of the entire catalogue and in introducing items that users have yet to encounter. As previously highlighted in the *APLT* analysis, it's important to recognize that *IBPR* and *WMF* stand out for their ability to feature *long-tail* items in their recommendations, thus enriching the diversity and reach of recommended content.

When analyzing user personas based on their privacy levels, it is observed from Fig. 4.9 that the highest values of entropy tend to be recorded for users with a high concern for their privacy, specifically *FF* users, with *PP* and *BB* following in order. Consequently, as the level of privacy protection for users increases, there is an enhanced likelihood of receiving recommendations of higher quality, where the representation of items across the spectrum is more equitable, albeit with few exceptions. This observation extends to the realm of novelty in recommendations, as corroborated by Fig. 4.10, where the most novel recommendations, denoted by the highest *Novelty* scores, are primarily seen for *FF* users, with *PP* and *BB* trailing respectively.

While the disparities in *Novelty* scores among user personas based on privacy levels are markedly pronounced for the ML dataset, they appear to be constrained for the more sparse and voluminous DB and Yelp datasets, a phenomenon attributable to the previously discussed characteristics of the datasets. Furthermore, although such variations in privacy level-based user personas may not be distinctly evident in the *LTC* outcomes for the DB and Yelp datasets, it is noticeable that users with the highest sensitivity towards privacy protection within the ML dataset are capable of receiving more refined recommendations in terms of catalogue coverage.

Ranking accuracy stands as a pivotal criterion for evaluating the efficacy of recommendation systems. Consequently, undertake a supplementary suite of experiments aimed at meticulously assessing how various recommendation algorithms exhibit differential performance across user personas categorized by their privacy levels, employing four pivotal accuracy-centric metrics (*Precision*, *Recall*, *F1-score*, and *nDCG*). In pursuit of this, delineate the achieved *Precision* and *Recall* scores of ten distinct algorithms across four

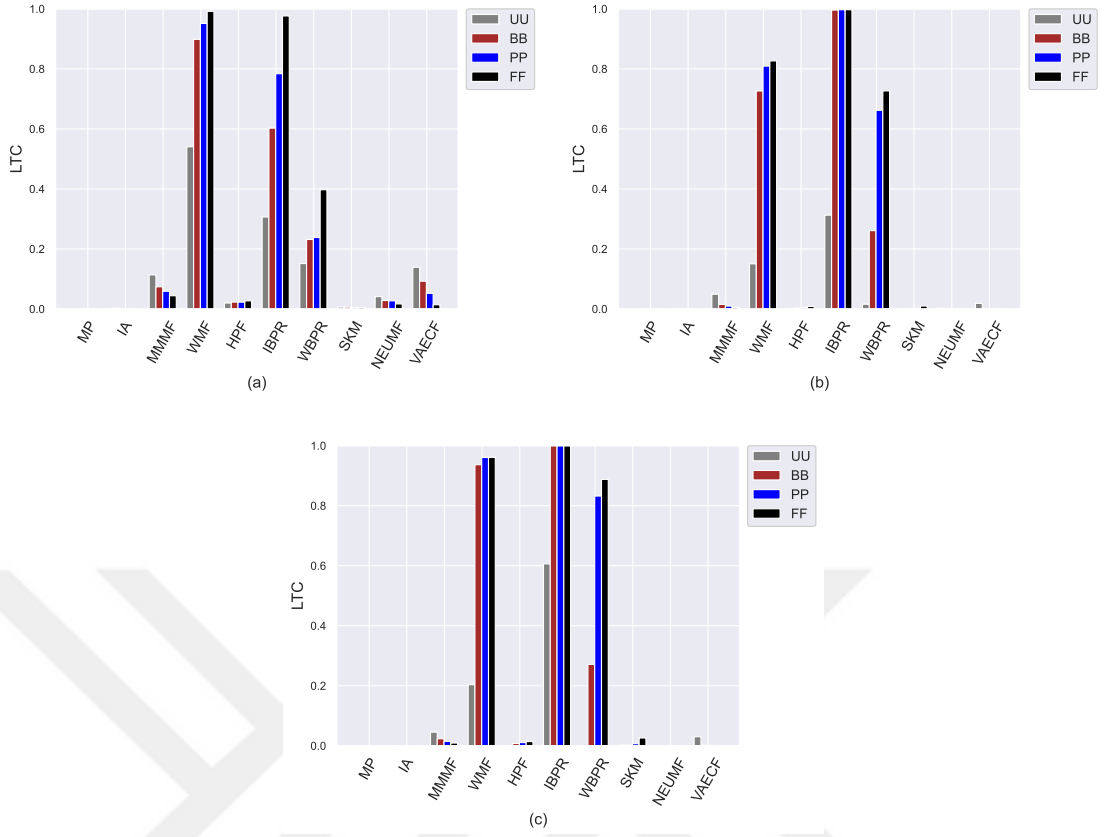


Figure 4.11. The *LTC* outcomes derived from the algorithms applied across user profiles categorized by four distinct levels of privacy within the (a) ML, (b) DB, and (c) Yelp datasets.

distinct user profiles in Fig. 4.12, alongside the *F1-score* and *nDCG* outcomes depicted in Fig. 4.13.

Given the recognized trade-off between precision and comprehensive recommendation quality, it's anticipated that algorithms excelling in aspects beyond accuracy might not exhibit optimal ranking precision or the converse. This paradox extends to user personas categorized by privacy levels, indicating that it's improbable for users to receive high-quality recommendations simultaneously in both accuracy and beyond-accuracy dimensions. The experimental findings focused on accuracy corroborate this discrepancy. For instance, algorithms that excel in beyond-accuracy quality, such as *IBPR*, typically manifest lower scores in *Precision*, *Recall*, *F1-score*, and *nDCG* compared to their counterparts, as evidenced by Figs. 4.12 and 4.13. Conversely, algorithms that fall short in beyond-accuracy performance, for example, those inspired by neural networks (*NE*-

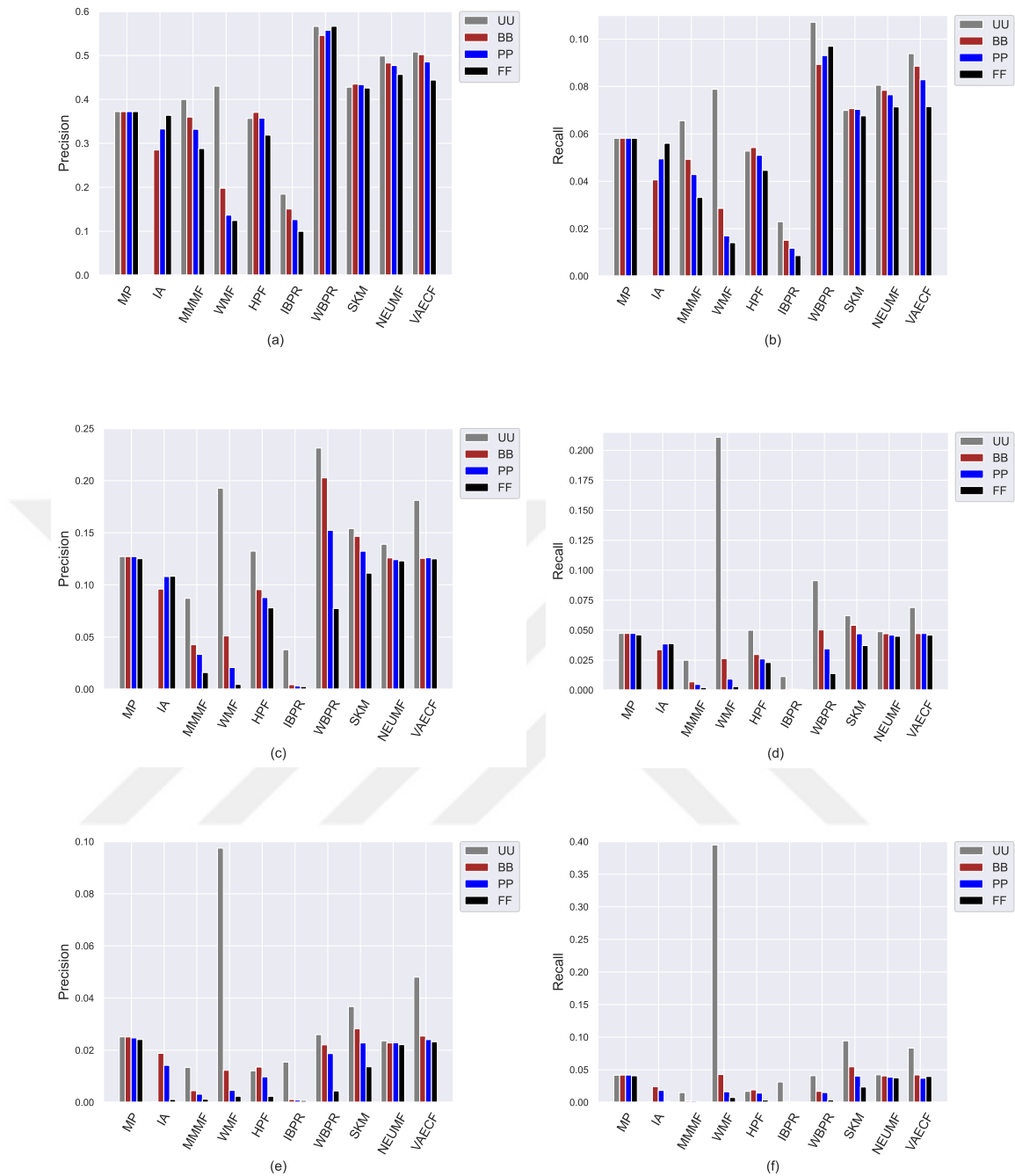


Figure 4.12. The derived *Precision* and *Recall* scores for the deployed algorithms across four distinct privacy level-based user profiles are detailed. *Precision* metrics are delineated in panels (a), (c), and (e) corresponding to the ML, DB, and Yelp datasets, respectively, whereas *Recall* metrics are depicted in panels (b), (d), and (f).

UMF and *VAECF*), demonstrate superior capability in delivering precise recommendations. Additionally, it is discernible that *WBPR* stands out as another algorithm profici-

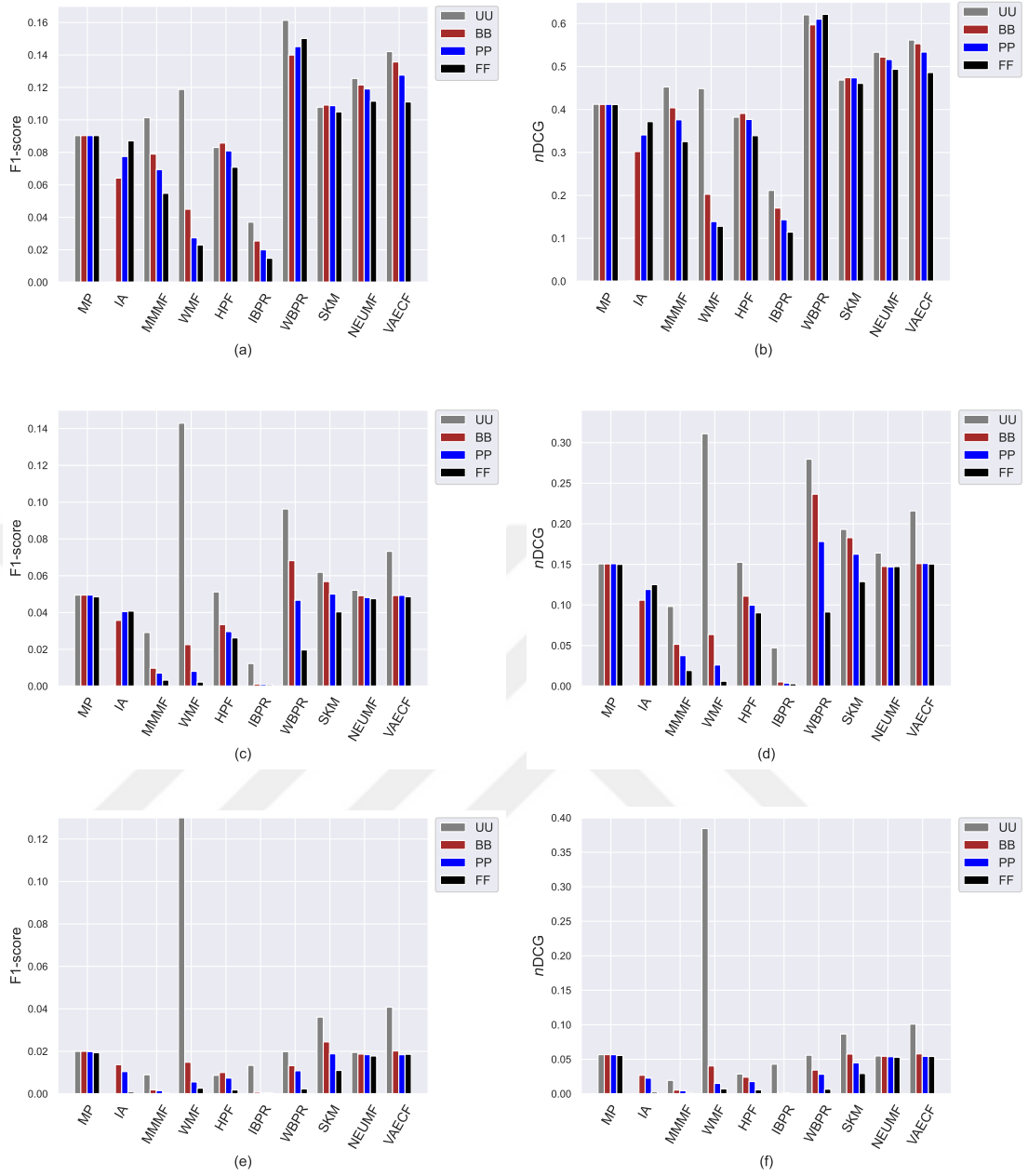


Figure 4.13. The derived $F1$ -score and $nDCG$ scores for the deployed algorithms across four distinct privacy level-based user profiles are detailed. $F1$ -score metrics are delineated in panels (a), (c), and (e) corresponding to the ML, DB, and Yelp datasets, respectively, whereas $nDCG$ metrics are depicted in panels (b), (d), and (f).

ent in furnishing accurate recommendations. Nevertheless, the results distinctly highlight the WMF algorithm's superiority in ranking precision when implemented without any privacy-preserving measures, specifically in the UU scenario. This revelation is particularly noteworthy given the prior acknowledgment of its prowess in beyond-accuracy

attributes. Hence, it is inferred that the *WMF* constitutes an advantageous choice for conventional recommendation contexts devoid of privacy safeguards.

As demonstrated by Figures 4.12 and 4.13, in the *UU* scenario, users typically receive more precise recommendations compared to those in privacy-enhanced settings across almost all evaluated metrics; this trend holds true except for the basic non-personalized *MP* and *IA* methods. In analyses of user personas based on privacy levels, the least favorable *Precision*, *Recall*, *F1-score*, and *nDCG* outcomes are observed for *FF*, followed sequentially by *PP* and *BB*. This pattern is consistent across all personalized CF techniques within the ML dataset; however, the disparity between such user personas based on privacy levels appears marginal for the *NEUMF* and *VAECF* algorithms within the DB and Yelp datasets. Hence, it can be inferred that an increase in the degree of privacy for users generally leads to a decline in the accuracy of the recommendations they receive. The primary rationale behind this observation is that algorithms are more adept at modeling individuals when processing their original profiles, resulting in highly accurate recommendations. However, their effectiveness substantially decreases when dealing with anonymized profiles. It's important to note that this outcome aligns with prior studies on PPCF methodologies (Bilge and Polat 2013).

4.4. Analyzing How Popularity-Debiasing Approaches Differently Perform for Privacy Level-based User Personas

This section delineates the empirical findings aimed at assessing the efficacy of traditional *popularity bias mitigation strategies* across user personas differentiated by their privacy levels. For this purpose, engage with three established debiasing techniques, namely *Aug*, *Mul*, and *xQd*, evaluating their impact on both privacy-enhanced user profiles (*BB*, *PP*, and *FF*) and unaltered original profiles (*UU*). The experimental analysis is carried out using the Yelp dataset, chosen for its larger size and sparser nature, reflecting common characteristics of real-world scenarios. Given the pronounced disparities among privacy-centric user personas particularly evident with the *HPF* algorithm, focus narrows to this algorithm during the recommendation phase to more accurately discern the debiasing interventions' effectiveness. Within these experimental confines, quantitatively presented the alterations in both beyond-accuracy metrics (*APRI*, *APLT*, *Entropy*, *Novelty*,

and *LTC*) and accuracy metrics (*Precision*, *Recall*, *F1-score*, and *nDCG*), as percentages, consequent to the application of *Aug*, *Mul*, or *xQd* in Figs. 4.14 and 4.15, respectively.

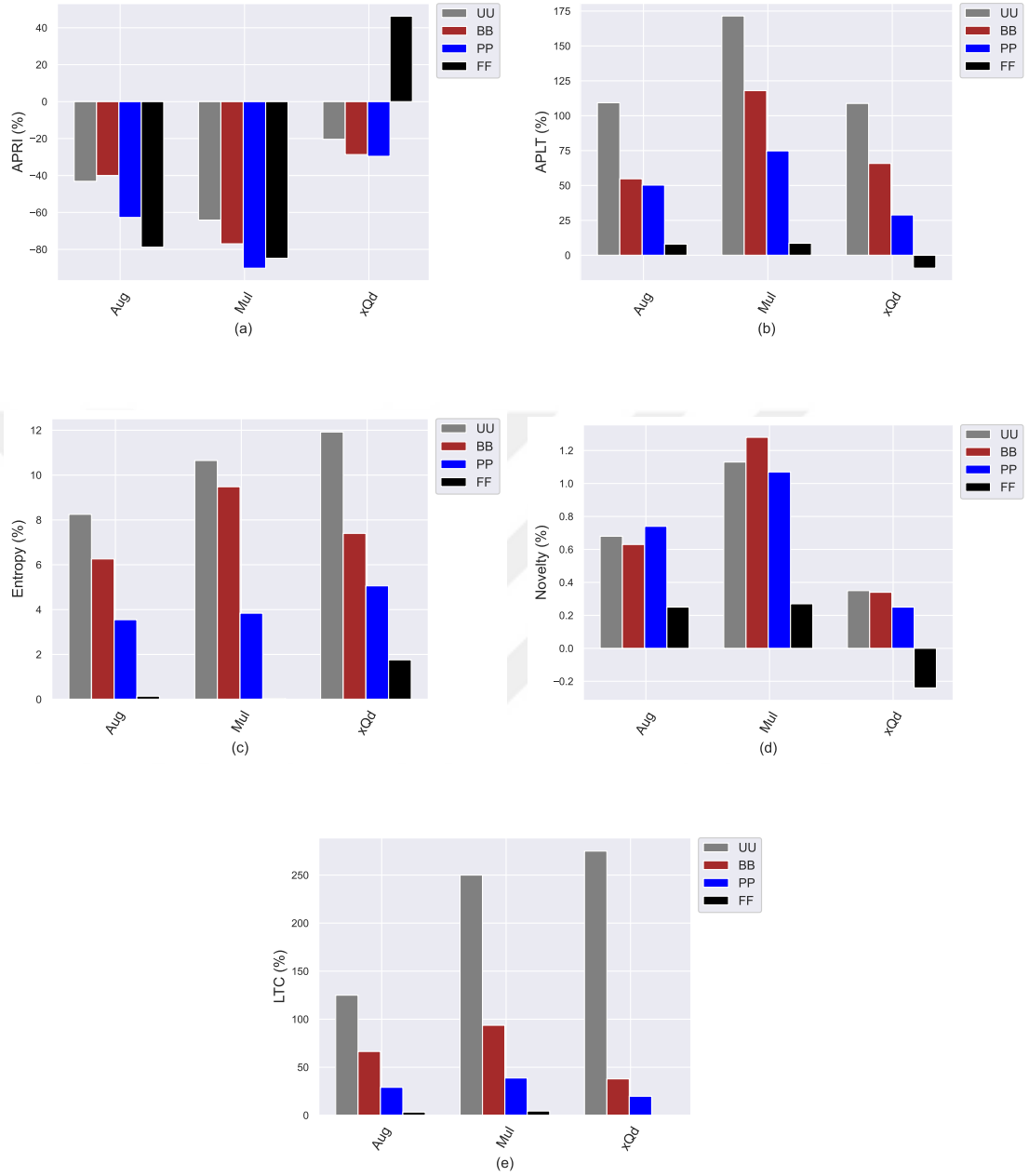


Figure 4.14. The enhancements or declines in (a) *APRI*, (b) *APLT*, (c) *Entropy*, (d) *Novelty*, and (e) *LTC* metrics observed upon the application of *Aug*, *Mul*, and *xQd* debiasing techniques for user personas classified by different privacy levels within the Yelp dataset.

As illustrated by Fig. 4.14, the most considerable reductions in *APRI* are found for *FF*, with subsequent decreases observed for *PP*, *BB*, and *UU*. Given that lower scores in this

metric signify recommendations that are less biased towards item popularity, it recommends that individuals who place a higher value on their privacy gain greater advantages from such debiasing techniques in countering popularity bias. Conversely, it's notably seen that these debiasing methods perform less effectively for users sensitive to privacy in terms of accuracy metrics. It is particularly noteworthy that the most pronounced decreases in ranking accuracy are seen for *FF*, with *PP*, *BB*, and *UU* following, as evidenced by the *Precision*, *Recall*, *F1-score*, and *nDCG* metrics in Fig. 4.15. This finding is intriguing since accuracy-related metrics typically evaluate the extent to which recommended items match with those rated in users' original profiles, leading to the expectation that recommending popular items would enhance overall accuracy performance.

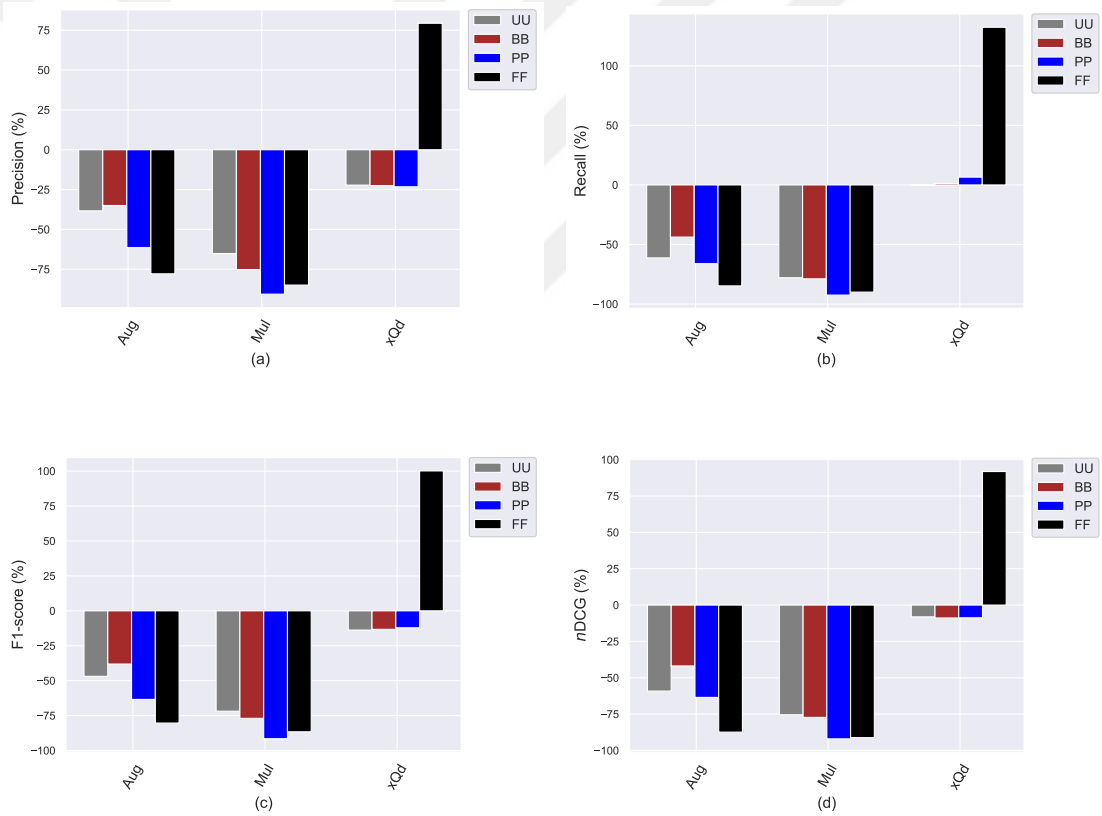


Figure 4.15. The enhancements or declines in (a) *Precision*, (b) *Recall*, (c) *F1-score*, and (d) *nDCG* metrics observed upon the application of *Aug*, *Mul*, and *xQd* debiasing techniques for user personas classified by different privacy levels within the Yelp dataset.

Contrastingly, the most substantial enhancements in the *APLT*, *Entropy*, *Novelty*, and *LTC* outcomes of recommendations are recorded for *UU*, with *BB*, *PP*, and *FF* following

in sequence. This pattern holds true across all implemented popularity-debiasing techniques, namely, *Aug*, *Mul*, and *xQd*. In essence, the popularity-debiasing methods under consideration prove more efficacious when factoring in the users' original profiles, and their proficiency in bolstering recommendation quality beyond accuracy diminishes as the privacy concern of users escalates. This discovery aligns with prior observations indicating that an enhancement in users' privacy levels leads to a noticeable improvement in the beyond-accuracy attributes of their recommendations. Consequently, the impact of the debiasing methods in use is curtailed for users with a heightened sense of privacy. Moreover, these privacy-conscious users also appear to be at a disadvantage in terms of accuracy, as the most significant reductions in ranking accuracy predominantly occur for *FF*, followed by *BB*, *PP*, and *UU*, as previously discussed.

One notable observation is the significant adverse effect observed when the highly regarded *popularity-debiasing* method, specifically the *xQd*, is applied to users with a high privacy preference, notably *FF* users. Remarkably, this leads to a deterioration in recommendation quality, as evidenced by negative improvements in *APLT* and *Novelty*, contrasted with positive outcomes for *APRI*. This key insight highlights that incorporating privacy-protective measures into user profiles can potentially neutralize the effectiveness of certain well-established popularity-debiasing strategies. Furthermore, empirical findings indicate that the efficacy of popularity-debiasing methods varies across user personas with different privacy levels, contingent upon the specific aspect of quality under consideration.

In this phase of the study, the *EquiRate* framework, which includes data perturbation and data obfuscation methods and beyond, is mentioned and the most appropriate method for the popularity bias problem is sought. In addition, the *FusionIndex* metric is also detailed for use in this and other studies. Therefore, the answers to RQ-3 and RQ-4 in the first chapter are sought.

4.5. The *EquiRate*: The Proposed Balanced Rating Injection Strategy for The Popularity-Debiasing Problem

The phenomenon of popularity bias in recommendation systems arises from the natural imbalances present in the distribution of data. In many instances, a minor subset of

items, typically those that are most popular or considered 'head' items, accumulate an inordinate share of user interactions and ratings. This phenomenon results in a lengthy 'tail' of items that are less popular, or 'tail' items, which are rarely recommended. Such a skewed distribution initiates a feedback loop, whereby items that are already popular receive increased visibility and become even more prevalent, leaving the less popular 'tail' items in obscurity. This cycle not only diminishes the diversity of recommendations available to users but also obstructs the exploration of novel or *niche* content, adversely impacting the overall user experience and the efficiency of the system (Boratto et al. 2021; Gulsoy et al. 2023).

The fundamental reason behind this bias is linked to the conventional design of recommendation systems. The majority of these systems' algorithms are constructed to enhance accuracy by relying on existing user-item interactions, inadvertently giving an edge to items that have accumulated a higher number of interactions. This creates a '*rich get richer*' scenario in which popular items receive an increasing number of recommendations and, consequently, more ratings, which in turn solidifies their popularity. Such a disproportionate focus on popular items can be counterproductive, as it restricts the diversity of content presented to users, often neglecting items that may be relevant but are less known.

To address this issue, introduce a novel strategy with method, *EquiRate*, which recalibrates the recommendation environment. *EquiRate* is deliberately engineered to counteract the impact of popularity bias through the introduction of synthetic ratings for items in the long-tail. This approach strives to equalize opportunities, enabling these lesser-known items to gain visibility and be considered for recommendation to users.

The *EquiRate* functions on the basis of enhancing the visibility of tail items within a dataset. By distinguishing items as '*popular*' and '*less popular*' according to the *Pareto principle*, it methodically tackles the issue of imbalance. The approach includes the strategic incorporation of synthetic ratings into these tail items, derived through meticulously crafted algorithms and formulas. These synthetic ratings are intentionally calculated to mirror realistic user preferences, ensuring the integrity of the recommender system is preserved. By meticulously injecting synthetic ratings, the *EquiRate* interrupts the feedback loop that favors the prominence of popular items. This action facilitates a broader and more equitable representation of items in recommendation lists, thus broadening the

spectrum of products, services, or content accessible to users, which might have remained undiscovered due to the prevalent focus on popular items.

In conclusion, the *EquiRate* offers a tactical remedy to counteract the data distribution skewness that fuels popularity bias in recommender systems. It achieves this by elevating the visibility and recommendation potential of less popular items, thereby cultivating a recommendation ecosystem that is more equitable, varied, and representative. This methodology not only enhances the user experience by expanding the spectrum of recommendations but also ensures a fairer platform for all items, independent of their original popularity. The paramount objective of the *EquiRate* is to support a recommendation environment where items are appraised and recommended based on their pertinence and merit, rather than predominantly on pre-existing popularity indicators.

To be more specific, the initial phase of the *EquiRate* method entails segmenting items into various popularity categories based on the *Pareto* principle. Let us define I as the set comprising all items, and R_i represents the count of ratings that item i has received. Subsequently, the items are arranged in a descending order according to R_i . The top 20% of these items are identified as 'popular' or 'head' items, signified by H , whereas the remaining 80% are designated as tail items, represented by T . To calculate the quantity of synthetic ratings to be incorporated into each tail item, the *EquiRate* adopts three separate strategies, known as **Synthetic Rating Injection (SRI)** strategies, which are elaborated upon as follows:

SRI Strategies

- **Overall Popularity Adjustment (OPA):** Employing this methodology, the popularity score (Pop_i) for each item i is first determined based on its received number of ratings and the average popularity score (Pop_{avg}) across all items. Subsequently, the *OPA* approach ascertains the quantity of synthetic ratings (n_i) to be allocated for each item in the tail group, as delineated in Eq. 4.9, with α serving as a scaling coefficient within the range of $[0, 1]$, which modulates the aggregate number of ratings to be introduced.

$$n_i = \alpha \times (Pop_{avg} - Pop_i) \quad (4.9)$$

- **Head Item Focused Adjustment (HIFA):** In a manner akin to the *OPA*, this approach initially calculates Pop_i for every item i . However, diverging from *OPA*, the *HIFA* methodology computes the average popularity score (Pop_{head_avg}) exclusively for head items. It then decides the amount of synthetic ratings to be distributed to each item in the tail category (n_i), as shown in Eq. 4.10, where α is once again a scaling factor within the $[0, 1]$ range.

$$n_i = \alpha \times (Pop_{head_avg} - Pop_i) \quad (4.10)$$

- **Threshold-Based Popularity Adjustment (TPA):** In alignment with preceding strategies, this method initially ascertains Pop_i for every item i within the catalog. Nonetheless, the *TPA* strategy introduces a distinct step by establishing a popularity threshold θ , derived from the minimal popularity found amongst head items. Subsequently, it ascertains the quantity of synthetic ratings to be assigned to each item classified under the tail category (n_i), as delineated in Eq. 4.11, where α persists as the scaling factor spanning the $[0, 1]$ interval.

$$n_i = \alpha \times (\theta - Pop_i) \quad (4.11)$$

After the calculation of the quantity of synthetic ratings to be injected, a subsequent essential phase involves the allocation of rating values to vacant positions within the profiles of items in the tail category. To tackle this challenge, implement a series of **Synthetic Rating Generation (SRG)** strategies depicted below. These methodologies are designed to introduce synthetic ratings in a way that preserves the coherence of item profiles while mitigating the impact of popularity bias:

- **Generalized Rating Value (GRV):** This approach involves computing the values of synthetic ratings by leveraging aggregate statistics from both item and user profiles. These statistics could encompass the average rating across all items, the average rating issued by a given user, or the average rating an item has received.
- **Predictive Rating Value (PRV):** In this method, adopt a more individualized strategy by generating synthetic ratings for items in the tail segment using recommen-

dation algorithms, such as those based on *collaborative filtering*. Initially, these algorithms are trained using the existing user-item interaction matrix and subsequently employed to generate user- and item-specific synthetic ratings.

- **Randomized Value Generation (RVG):** This strategy is influenced by privacy-preserving collaborative filtering methods (Polat and Du 2005b; Bilge and Polat 2013) designed to protect user privacy within recommender systems. Employ *randomized perturbation* (Gulsoy et al. 2023), a widely adopted technique for randomly altering each actual rating. Consequently, rating vectors generated at random will fill the vacant cells in the actual rating vectors for items in the tail. Consistent with the initial approach of *randomized perturbation*, these vectors are produced using either a uniform or a normal distribution. For the uniform distribution, random numbers are generated within the interval of $[-\sqrt{3}\sigma, \sqrt{3}\sigma]$, while for the normal distribution, random numbers adhere to the $N(\mu, \sigma_{\max}^2)$ distribution. It is important to note that the mean of the randomly generated numbers in both distributions is zero.

Algorithm 2: Procedures of the proposed *EquiRate* method

Data: User-Item Rating Matrix R , Scaling Factor α
Result: Balanced User-Item Rating Matrix R'
 Classify items into H and T using Pareto Principle [ref];
foreach *strategy* (i.e., *OPA*, *HIFA*, or *TPA*) **in** SRI **do**
 Calculate the necessary number of synthetic ratings for each item in T ;
 foreach *item* i **in** T **do**
 for j **from** 1 **to** *Number of Synthetic Ratings* **do**
 Select a rating value calculation method from SRG , (i.e., *GRV*, *PRV*, or *RVG*);
 Generate a synthetic rating v ;
 Append v to R' for item i ;
 end for
 end foreach
return R' ;

Figure 4.16. Procedures of the proposed the *EquiRate* method

In summary, the *EquiRate* method for mitigating popularity bias is detailed in figure 4.16. This methodology can be applied in nine different variations by selecting one of three distinct strategies from both the SRI and SRG categories. Subsequent experiments aim to ascertain the most effective combination for reducing popularity bias. Upon identifying the optimal *EquiRate* variant, it can serve to pre-process a dataset for utilization with

any recommendation algorithm, including those reliant on collaborative filtering techniques. This pre-processing step yields a balanced rating dataset (represented as R'), which, when used in lieu of the original dataset for recommendation algorithms, is anticipated to enhance the visibility of less popular (tail) items in the recommendations. Thus, it potentially elevates beyond-accuracy attributes such as diversity, novelty, and fairness within the recommendations. Furthermore, provide a schematic overview of the general framework of the proposed *EquiRate* method for counteracting popularity bias in Fig. 4.17.

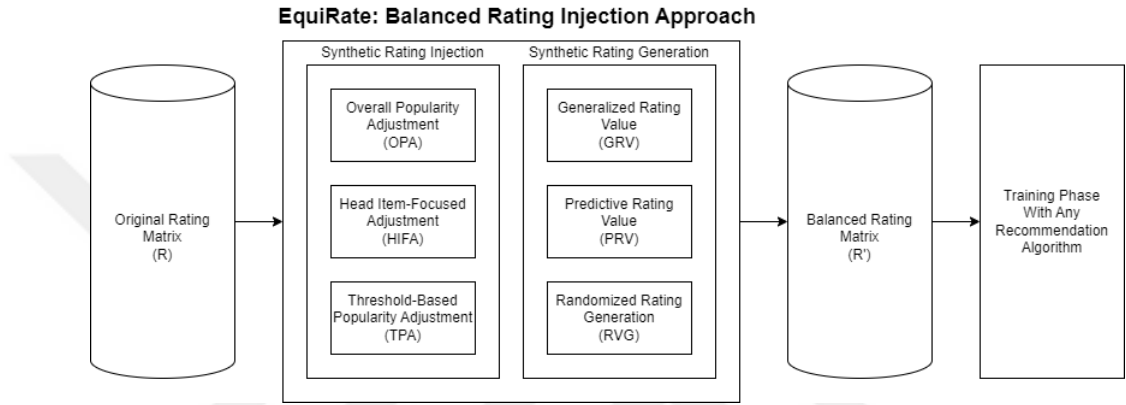


Figure 4.17. General structure of the proposed *EquiRate* approach

4.6. Experimental Studies

This section elucidates the datasets employed, the parameter configuration for the *EquiRate* approach, and the benchmark *popularity debiasing* methods for comparative analysis. Moreover, it outlines an innovative evaluation metric that simultaneously addresses the *accuracy* and *beyond-accuracy* dimensions of recommendation systems. Furthermore, the section delineates the outcomes of the experimental investigations conducted.

4.6.1. Parameter Setup

As outlined in Section 4.5, the experimentation involves two pivotal aspects: firstly, the allocation of synthetic ratings to specific items employing any *SRI technique*, and secondly, the computation of these ratings' values using any *SRG strategy*.

Nonetheless, increasing the variety of experimental scenarios enhances the exploration of parameter configurations where the proposed approach might excel. Consequently,

the initial nine scenarios, utilizing a blend of three *Synthetic Rating Injection (SRI)* techniques (namely, *OPA*, *HIFA*, and *TPA*) and three *Synthetic Rating Generation (SRG)* methods (namely, *GRV*, *PRV*, and *RVG*), are expanded into 156 distinct scenarios through the variation of parameter values. The α parameter, governing the density of ratings introduced by SRI methods, is set across eight different levels ranging from 0 to 1 (i.e., 0.1, 0.25, 0.33, 0.50, 0.66, 0.75, 0.90, and 1.00). Additionally, the formulation of synthetic ratings via SRG methods incorporates three diverse methodologies. Specifically, the GRV approach is evaluated using three sub-strategies: the average ratings of a given user (u_{avg}), the average rating of a specific item (i_{avg}), and the overall average rating within the dataset (all_{avg}). Meanwhile, the PRV method integrates the VAECF algorithm, recognized for its precision as one of the foremost *collaborative filtering* techniques in contemporary research (Liang et al. 2018). Lastly, the RVG method is assessed under three different scenarios by modulating the volume of injected ratings through varying privacy settings: *Low*, *Middle (Mid)*, and *High*. In this context, for the low scenario, σ and β are opted as 2 and 5, respectively; for the middle scenario, they are set at 3 and 10; and for the high scenario, at 4 and 25. It's also noteworthy that due to the negligible discrepancies between uniform and normal distributions observed in the experiments concerning the RVG method, only results derived from the Normal distribution are presented.

In conclusion, the integration of three *SRI* methods, eight distinct α parameters, and seven *SRG* methods culminates in a total of 156 experimental scenarios. Nevertheless, for the sake of simplicity, this study will only detail the performance of the most effective variants of the *EquiRate* method in the ensuing section. Specifically highlight results with α values set at 0.1, 0.5, and 0.9 to more thoroughly assess their impact on the effectiveness of the final recommendation outcomes. To ensure the reliability of these findings, each variant is executed five times, and the average of these iterations is calculated to mitigate the influence of randomness and enhance the consistency of the results. Figure 4.18 illustrates the comprehensive parameter adjustments tested for the *EquiRate* methodology. In the culmination of research, apply the VAECF algorithm to a dataset refined by *EquiRate* adaptations, with the goal of producing top- N recommendation lists for each user, selecting N as 10. It is also pertinent to mention that the implementations of the VAECF algorithm were performed using the Cornac framework, a modern Python-based

platform⁵ (Salah et al. 2020).

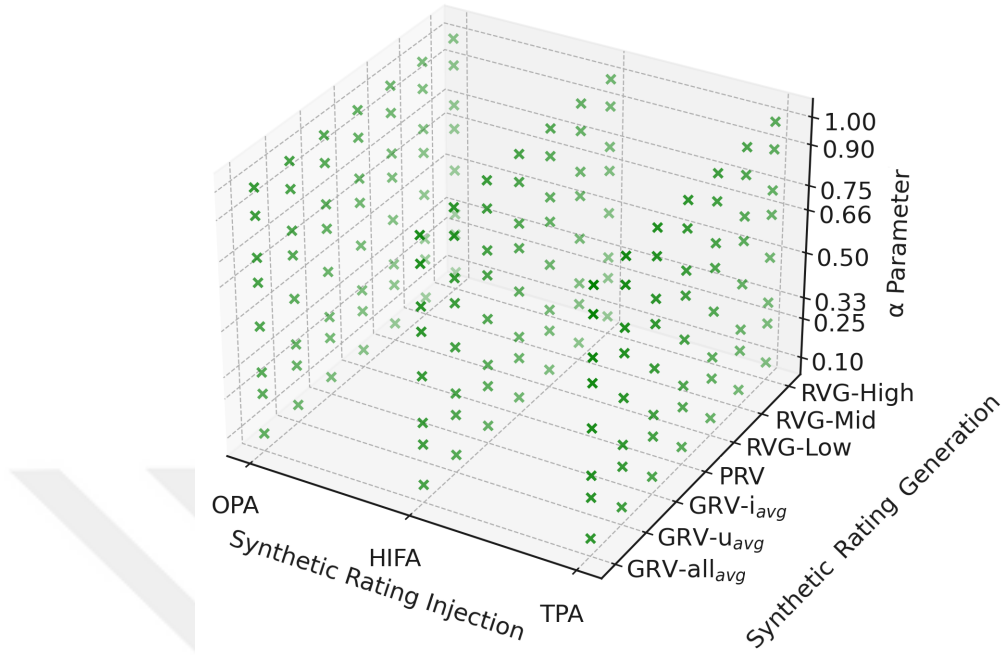


Figure 4.18. All parameter adjustments tested for the *EquiRate* are documented. The green dots symbolically illustrate each variant of the *EquiRate*.

4.6.2. Evaluation Metrics

To conduct a comprehensive evaluation of the *EquiRate* approach proposed and adopted a set of standards for assessing recommendation quality in this experiments, mirroring the approach of a similar study (Yalcin and Bilge 2022). In evaluating recommendation accuracy, rely on *nDCG*, the most acknowledged and paramount accuracy metric. Moreover, incorporate four supplementary metrics to gauge the quality of recommendations beyond accuracy. These include *Average Percentage of Long Tail Items (APLT)* (Abdollahpouri et al. 2021; Abdollahpouri 2020), *Entropy* (Elahi et al. 2021), *Novelty* (Yalcin and Bilge 2022), and *Long-Tail Coverage (LTC)* (Abdollahpouri 2020).

Crucially, these metrics assess recommendation lists through diverse lenses, leading to varying performances of popularity-debiasing strategies across these metrics. In essence,

⁵<https://cornac.preferred.ai/>

a debiasing technique might excel in one metric but fall short in capturing another dimension of recommendation quality. This complexity presents a significant challenge in determining the superiority of one debiasing method over others when considering multiple facets of recommendation quality simultaneously. To address this challenge, one of the principal contributions of this research is the introduction of a new evaluation metric, the *FusionIndex*, designed to harmonize these metrics. This section outlines the established metrics and provides an in-depth explanation of the newly proposed *FusionIndex*.

4.6.3. The *FusionIndex*: A Highly Useful Holistic Evaluation Metric

In the field of recommendation systems, it is of paramount importance to accurately evaluate the efficacy of various algorithms. Conventional metrics have typically concentrated on either accuracy or measures beyond accuracy, not fully capturing the comprehensive performance of a system. To bridge this gap, propose the *FusionIndex*, an innovative metric that integrates considerations of accuracy with factors beyond accuracy.

The *FusionIndex* initiates its process by calculating the values of several critical metrics for a top- N recommendation list. At the heart of the *FusionIndex* is its distinctive method of integration. Rather than considering these metrics in isolation, the *FusionIndex* employs the harmonic mean to combine specific pairs, highlighting the synergy between accuracy ($nDCG$) and various beyond-accuracy metrics. The selection of the harmonic mean is due to its adeptness at addressing skewed distributions, ensuring that the evaluation is not disproportionately influenced by either the *accuracy* or the *beyond-accuracy* metrics. The pairs and their associated harmonic means are determined according to Eqs. [4.12] - [4.15]. Specifically, the harmonic mean between $nDCG$ and the metrics *APLT*, *Novelty*, *LTC*, and *Entropy* is calculated using Eqs. [4.12], [4.13], [4.14], and [4.15], respectively.

$$HM_{nDCG-A} = \frac{2}{\left(\frac{1}{nDCG} + \frac{1}{APLT}\right)} \quad (4.12)$$

$$HM_{nDCG-N} = \frac{2}{\left(\frac{1}{nDCG} + \frac{1}{Novelty}\right)} \quad (4.13)$$

$$HM_{nDCG-L} = \frac{2}{\left(\frac{1}{nDCG} + \frac{1}{LTC}\right)} \quad (4.14)$$

$$HM_{nDCG-E} = \frac{2}{\left(\frac{1}{nDCG} + \frac{1}{Entropy}\right)} \quad (4.15)$$

Finally, the *FusionIndex* computes the arithmetic mean of these four harmonic means to obtain a unified and comprehensive score for the recommendation list, as defined in Eq. 4.16.

$$FusionIndex = \frac{HM_{nDCG-A} + HM_{nDCG-N} + HM_{nDCG-L} + HM_{nDCG-E}}{4} \quad (4.16)$$

This method guarantees a balanced evaluation of both accuracy and beyond-accuracy dimensions. By employing the harmonic mean, the *FusionIndex* reduces the likelihood that one dimension will unduly influence the overall analysis, thus offering a more detailed and thorough assessment of recommendation systems. In essence, the *FusionIndex* introduces an innovative and equitable approach for assessing recommendation systems, recognizing the significance of both precision and variety in recommendations. This index aids in a deeper comprehension of the performance of recommendation algorithms, facilitating the development of more sophisticated and consumer-focused recommendation systems.

4.6.4. Experiment Results

This section delineates the outcomes of the conducted experiments aimed at thoroughly assessing the introduced *EquiRate* approach and juxtaposing it against established benchmarks in popularity bias mitigation methods.

4.6.5. In-depth Analysis of The Proposed *EquiRate* Method

This section elucidates the enhancement rates observed in the *FusionIndex* metric upon the application of various *EquiRate* versions, aiming to discern the top-performing configurations and to examine the influence of *SRI*, *SRG*, and α parameters on the *EquiRate* strategy. Utilize multiple *EquiRate* variants in these analyses by varying *SRI*, *SRG*, and α parameters. Figures 4.19-4.21 illustrate the optimal *FusionIndex* enhancements achieved by the *EquiRate* versions across MLM, DB, and Yelp datasets, respectively. In

these figures, each *EquiRate* variant is denoted by the specific SRI, α , and SRG approaches taken into account. It's important to highlight that these experiments exclusively employ the *FusionIndex* metric, foregoing individual metrics like *nDCG*, *LTC*, *Entropy*, or *APLT*, due to its capacity to aptly and simultaneously evaluate the recommendations on the basis of both *accuracy* and *beyond-accuracy* measures, as elaborated in Section 4.6.3.

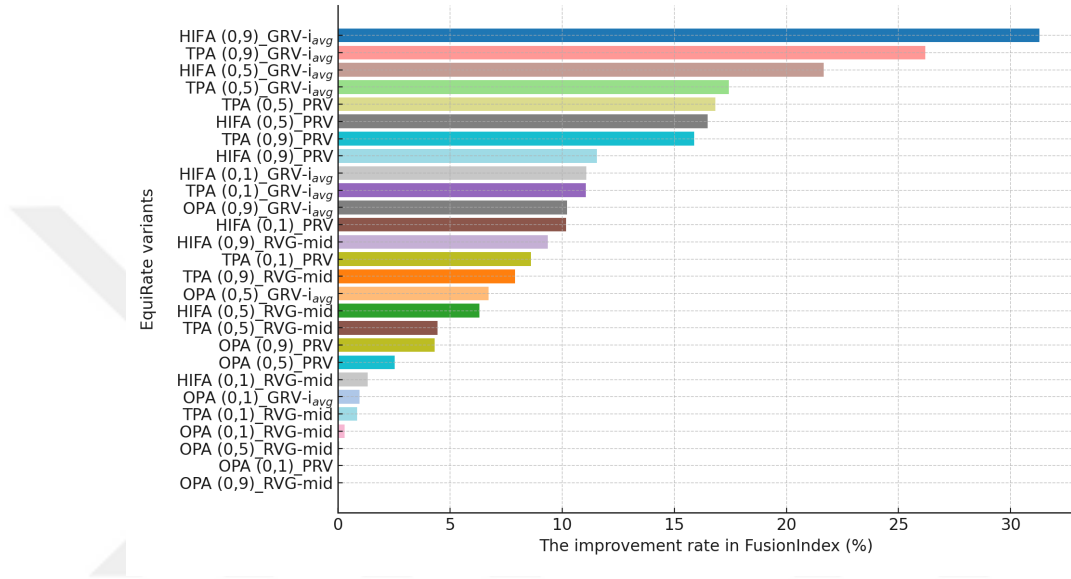


Figure 4.19. The enhancement in the *FusionIndex* percentage through various implementations of the *EquiRate* methodology in the MLM dataset is documented. In this context, the different versions of *EquiRate* are identified based on the specific SRI, α , and SRG strategies that were employed.

As illustrated in Fig. 4.19, within the MLM dataset, the *HIFA (0.9)_GRV-i_avg* version of *EquiRate* demonstrated superior performance, achieving a 31.29% enhancement in the *FusionIndex* score. This improvement underscores the efficiency of the HIFA methodology and a higher α value in diminishing the bias towards popular items, effectively adjusting the skew towards more popular products and enhancing the representation of *niche* content. Following closely, the *TPA (0.9)_GRV-i_avg* variant, with a 26.20% enhancement, evidences the effectiveness of a threshold-based method in bridging the divide between highly popular and less known items. The elevated α value facilitates a corrective measure against the popularity bias. Additionally, the *HIFA (0.5)_GRV-i_avg* model,

with a 21.65% boost, exemplifies a moderated yet significant improvement by employing a lower α value, crucially contributing to the refinement of both accuracy and diversity within recommender systems.

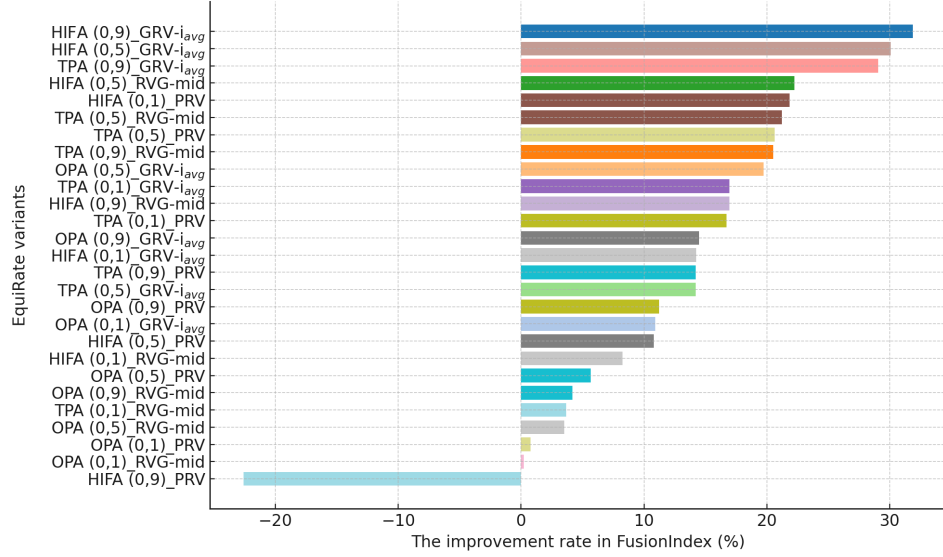


Figure 4.20. The enhancement in the *FusionIndex* percentage through various implementations of the *EquiRate* methodology in the DB dataset is documented. In this context, the different versions of *EquiRate* are identified based on the specific SRI, α , and SRG strategies that were employed.

Conversely, in the context of the DB dataset, the effectiveness of the *HIFA (0.9)_GRV- i_{avg}* variant stands out significantly, as evidenced by Fig. 4.20. This approach demonstrates the potential of a high α value in sparse data environments to generate impactful outcomes. It promotes the exposure of less popular items in recommendation lists by incorporating synthetic ratings, enhancing diversity. Both the *HIFA (0.5)_GRV- i_{avg}* and *TPA (0.9)_GRV- i_{avg}* versions also achieve commendable results, with their respective medium and high α values, effectively mitigating popularity bias within the DB dataset. It is also noteworthy that these findings are in close alignment with the patterns observed in the MLM dataset, indicating a consistent effectiveness across different datasets.

In conclusion, the results presented for the Yelp dataset, as shown in Fig. 4.21, highlight the superior performance of the *TPA (0.9)_PRV* variant, with an impressive improvement of 58.28%, illustrating the method's efficacy in handling large and sparse data

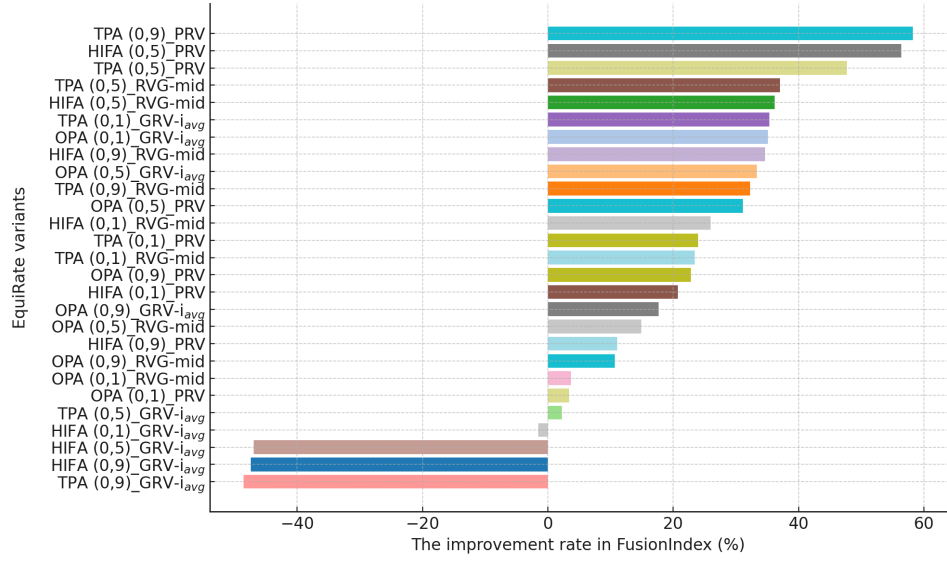


Figure 4.21. The enhancement in the *FusionIndex* percentage through various implementations of the *EquiRate* methodology in the Yelp dataset is documented. In this context, the different versions of *EquiRate* are identified based on the specific SRI, α , and SRG strategies that were employed.

environments. This underscores the successful integration of the PRV approach with personalized recommendation systems, offering a potent solution for navigating complex data landscapes. Both the *HIFA (0.5)_PRV* and *TPA (0.5)_PRV* variants also mark considerable progress with their medium α values, affirming their capability to mitigate popularity bias and enrich the diversity within the Yelp dataset’s recommendations. Furthermore, it is observed that the optimal *EquiRate* variants for Yelp slightly deviate from those identified in the MLM and DB datasets, recommending that *EquiRate* variants may exhibit varied performances based on the specific dataset in use. Nonetheless, it is evident that significant enhancements in the *FusionIndex* metric can be realized through meticulous parameter tuning of the *EquiRate* across any dataset.

The analysis of the *EquiRate* method’s performance across the MLM, DB, and Yelp datasets demonstrates its proficiency in diminishing popularity bias within recommender systems and augmenting the diversity of the recommendations provided. The findings from each dataset underscore the effective deployment of *EquiRate*’s SRI and SRG tactics, with a particular emphasis on the influence of the α scaling factor. Specifically, wit-

hin the MLM dataset, leveraging high α values in conjunction with the HIFA and TPA strategies presents a potent method for addressing popularity imbalances. This approach markedly boosts the visibility of *niche* or infrequently evaluated items in the recommendation lists, thereby increasing user satisfaction and the overall efficacy of the system. For the DB dataset, integrating high α values with the HIFA strategy has been proven to rectify popularity disparities in sparse data configurations effectively. In the context of the Yelp dataset, the amalgamation of the TPA strategy with the PRV method within extensive and intricate data structures significantly improves both the precision and diversity of the recommendations, enhancing the recommender system’s performance.

These findings clarift the distinct impacts of the *EquiRate*’s SRI strategies (OPA, HIFA, and TPA) and SRG strategies (GRV, PRV, and RVG) at varying levels of the α parameter. Specifically, elevated α values adopt a more assertive stance in addressing the issue of popularity bias, while intermediate α values yield more equitable effects. Personalization techniques such as PRV significantly improve the precision of recommender systems by more accurately mirroring user predilections, whereas strategies like GRV and RVG amplify the exposure of diverse and yet-to-be-discovered content.

Additionally, the robustness of the experimental findings is enhanced through multiple iterations and averaging these repetitions. This methodology ensures the stability and dependability of the results, offering a more faithful representation of the method’s efficacy. In summary, the advancements made by different variants of the *EquiRate* method in diminishing popularity bias and improving the diversity and precision of recommender systems across varied datasets are considerable. The capacity of the method to adapt to the distinct characteristics of each dataset and the pivotal role of the α scaling factor in achieving these enhancements are clear. This underscores the capacity of the *EquiRate* to refine recommender systems and augment the user experience.

4.6.6. Comparison of The *EquiRate* With Existing Popularity-Debiasing Methods

In this segment, represent the outcomes of supplementary experiments in which sorted the top-performing *EquiRate* variants against a suite of benchmark methods aimed at mitigating popularity bias. The benchmarks, namely *AUG*, *xQuad*, *PAW*, *MUL*, and *LNSF*, are thoroughly detailed in Section 2.4. of this paper. For these analyses, focus on the

premier ten variants of the *EquiRate* for each dataset, identified through the experiments conducted in the preceding section.

While the introduced *FusionIndex* serves as a sophisticated measure that evaluates accuracy and beyond-accuracy metrics simultaneously, diverging from prior experiments, this segment delineates both the *FusionIndex* and the individual outcomes for accuracy (i.e., *nDCG*) and beyond-accuracy metrics (i.e., *APLT*, *Novelty*, *LTC*, and *Entropy*) in these experiments to furnish a more comprehensive overview. In alignment with this approach, initially disclose the accuracy and beyond-accuracy outcomes of the recommendation lists generated through *EquiRate* variants and other benchmark methods for the MLM, DB, and Yelp datasets, delineated in Tables [4.6-4.8], correspondingly. It's important to note that the outcomes of the original recommendation algorithm prior to the application of any popularity-debiasing approach are also presented in these tables; this allows for an assessment of the extent to which popularity-debiasing methods can alleviate this concern in the final recommendations. For a holistic evaluation, the improvement ratios in the *FusionIndex* via the top ten performing *EquiRate* variants and established popularity debiasing methods for the MLM, DB, and Yelp datasets are also exhibited in Figs. [4.22-4.24], respectively.

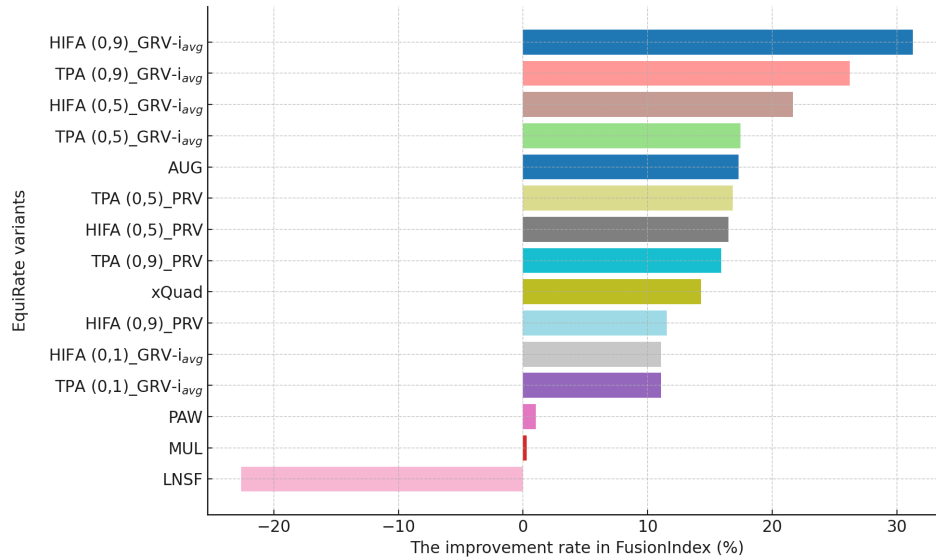


Figure 4.22. The enhancement in the *FusionIndex* scores following the application of popularity-debiasing techniques for the MLM dataset.

Table 4.6. The assessment of the *EquiRate* variants in comparison with various popularity-debiasing techniques based on *nDCG*, *APLT*, *Novelty*, *LTC*, and *Entropy* metrics specifically for the MLM dataset.

<i>Debiasing Method</i>	<i>SRI</i>	α	<i>SRG</i>	<i>Approach</i>	<i>nDCG</i>	<i>APLT</i>	<i>Novelty</i>	<i>LTC</i>	<i>Entropy</i>
<i>Original</i>					0.620	0.295	0.296	0.150	0.636
<i>EquiRate Variants</i>	HIFA	0.9	GRV	i_{avg}	0.556	0.417	0.367	0.744	0.747
	TPA	0.9	GRV	i_{avg}	0.597	0.383	0.323	0.550	0.714
	HIFA	0.5	GRV	i_{avg}	0.608	0.366	0.311	0.445	0.699
	TPA	0.5	GRV	i_{avg}	0.610	0.367	0.310	0.350	0.689
	TPA	0.5	PRV		0.534	0.413	0.400	0.350	0.702
	HIFA	0.5	PRV		0.492	0.434	0.447	0.388	0.715
	TPA	0.9	PRV		0.459	0.483	0.484	0.407	0.728
	HIFA	0.9	PRV		0.401	0.536	0.548	0.443	0.743
	HIFA	0.1	GRV	i_{avg}	0.597	0.364	0.327	0.246	0.671
	TPA	0.1	GRV	i_{avg}	0.595	0.368	0.329	0.242	0.674
<i>AUG</i>					0.452	0.827	0.453	0.285	0.783
<i>xQuad</i>					0.568	0.630	0.338	0.171	0.691
<i>PAW</i>					0.620	0.302	0.296	0.158	0.639
<i>MUL</i>					0.309	0.994	0.607	0.396	0.812
<i>LNSF</i>					0.356	0.831	0.984	0.936	0.916

According to the data in Table 4.6 for the MLM dataset, the *nDCG* value of the original algorithm prior to the implementation of any debiasing techniques stands at 0.620. Regrettably, all variants of *EquiRate* record lower *nDCG* scores, signifying a minor reduction in precision attributed to these interventions, which is in line with anticipations. This finding highlights the inherent balance between maintaining accuracy and addressing beyond-accuracy metrics in recommendations while attempting to neutralize popularity bias. Nonetheless, the relatively elevated *nDCG* levels imply that the recommendations may remain relevant and appealing to users. Conversely, compared to the *EquiRate* variants, other methods designed to counteract popularity bias, particularly *MUL*, tend to exhibit greater reductions in *nDCG* performance. This suggests that the slight accuracy compromises introduced by *EquiRate* approaches are deemed tolerable.

In the context of *APLT*, where the baseline value is 0.295, the application of high α

values in the *EquiRate* strategy (for instance, achieving 0.536 when SRI is set to HIFA, $\alpha = 0.9$, and SRG is defined as PRV) results in notable increments, as illustrated in Table 4.6. This showcases the propensity of enhanced recommender systems through *EquiRate* framework to promote recommendations of less popular, *niche* items, thereby broadening the diversity of content available to users. Similarly, the metric of *Novelty*, initially measured at 0.296, has seen improvements across different *EquiRate* configurations. Notably, the aforementioned *EquiRate* setup has elevated the *Novelty* score to 0.548, significantly enriching the novelty aspect of the recommendation lists and indicating a shift towards offering more unique and innovative items. In the case of *LTC*, with a starting point of 0.150, implementations of high α values within the *EquiRate* framework demonstrate substantial growth (e.g., reaching 0.744 when SRI is HIFA, $\alpha = 0.9$, and SRG is $\text{GRV-}i_{avg}$), which signifies an increased inclusion of *niche*, *long-tail* items in the recommendations, enhancing the exposure of lesser-known items. Furthermore, the *Entropy* metric, initially at 0.636, generally sees an uptrend with the adoption of *EquiRate* approaches. Variants with high α values, such as 0.747 in the case mentioned above, show elevated *Entropy* levels, indicating that the recommendations have grown more diverse, presenting users with a wider array of choices. Consequently, the various configurations of *EquiRate*, especially those with high α values, effectively mitigate popularity bias and enhance diversity within the MLM dataset. These findings suggest the potential for an improved balance between accuracy and diversity in recommendation systems.

Additionally, while some existing popularity-debiasing techniques such as *MUL* and *LNSF* may surpass *EquiRate* variants in terms of beyond-accuracy metrics, they significantly compromise the *nDCG* of recommendations when compared to *EquiRate* variants, as demonstrated in Table 4.6. This reduction in *nDCG* is counterproductive to the primary objective in addressing popularity bias, which is to enhance the beyond-accuracy aspects of recommendations without substantially sacrificing accuracy. However, evaluating the performance of debiasing methods based solely on individual metrics may not provide a holistic understanding of which approach yields the best overall outcomes. Hence, also offer insights into the overall efficacy of different *EquiRate* variants and other debiasing strategies within the MLM dataset through the improvement ratios presented in the *FusionIndex*, as depicted in Fig. 4.22.

As depicted in Fig. 4.22, variants of *EquiRate*, particularly those with elevated α values like HIFA and TPA, have attained the highest *FusionIndex* scores. Specifically, the HIFA (0.9)_GRV- i_{avg} variant reached the pinnacle with a score of 31.286%, while the TPA (0.9)_GRV- i_{avg} variant also showcased a substantial performance, registering a score of 26.203%. These outcomes underscore the efficacy of such variants in bolstering diversity and novelty within recommendation lists. This evidence further highlights the paramount importance of high α values in mitigating popularity bias and enriching diversity. Amongst other debiasing techniques, *AUG* and *xQuad* have demonstrated promising results, with *AUG* exhibiting a commendable performance at 17.276% and *xQuad* achieving a moderate performance at 14.278%. Nonetheless, these findings imply that the various *EquiRate* variants surpass these benchmark methodologies in overall effectiveness for the MLM dataset.

Table 4.7. The assessment of the *EquiRate* variants in comparison with various popularity-debiasing techniques based on *nDCG*, *APLT*, *Novelty*, *LTC*, and *Entropy* metrics specifically for the DB dataset.

<i>Debiasing Method</i>	<i>SRI</i>	α	<i>SRG</i>	<i>Approach</i>	<i>nDCG</i>	<i>APLT</i>	<i>Novelty</i>	<i>LTC</i>	<i>Entropy</i>
<i>Original</i>					0.283	0.079	0.715	0.016	0.457
<i>EquiRate Variants</i>	HIFA	0.9	GRV	i_{avg}	0.189	0.664	0.796	0.658	0.808
	HIFA	0.5	GRV	i_{avg}	0.213	0.613	0.771	0.216	0.637
	TPA	0.9	GRV	i_{avg}	0.216	0.588	0.769	0.190	0.612
	HIFA	0.5	RVG	mid	0.259	0.477	0.729	0.025	0.528
	HIFA	0.1	PRV		0.271	0.344	0.730	0.026	0.545
	TPA	0.5	RVG	mid	0.279	0.336	0.720	0.018	0.511
	TPA	0.5	PRV		0.229	0.534	0.761	0.059	0.603
	TPA	0.9	RVG	mid	0.255	0.476	0.734	0.023	0.523
	OPA	0.5	GRV	i_{avg}	0.251	0.490	0.737	0.023	0.529
	TPA	0.1	GRV	i_{avg}	0.230	0.560	0.754	0.035	0.539
<i>xQuad</i>					0.227	0.581	0.767	0.027	0.515
<i>AUG</i>					0.224	0.454	0.771	0.039	0.581
<i>PAW</i>					0.280	0.095	0.718	0.019	0.465
<i>MUL</i>					0.139	0.791	0.846	0.060	0.606
<i>LNSF</i>					0.208	0.774	0.793	0.048	0.573

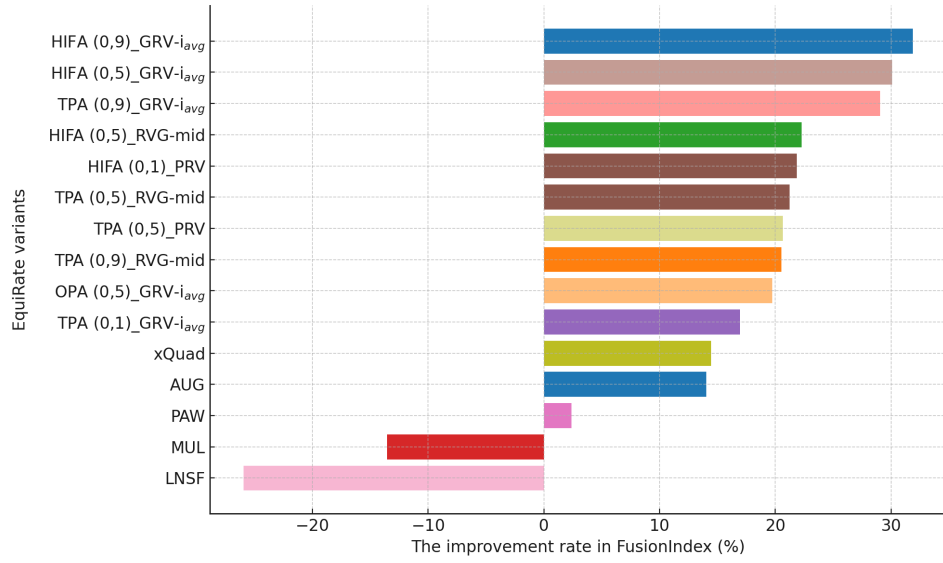


Figure 4.23. The enhancement in the *FusionIndex* scores following the application of popularity-debiasing techniques for the DB dataset.

Upon examining the outcomes presented in Table 4.7 for the DB dataset, it is observed that for $nDCG$, starting from an initial value of 0.283, all variants of *EquiRate* have resulted in reduced accuracy scores, echoing the patterns seen in the MLM dataset. Nevertheless, strategies employing high α values, such as TPA (for instance, achieving 0.279 when SRI is set to TPA, $\alpha = 0.5$, and SRG to RVG_{mid}) and HIFA (for example, securing 0.271 when SRI is HIFA, $\alpha = 0.1$, and SRG is PRV), have only marginally decreased $nDCG$ values, thereby delivering recommendations with a satisfactory level of accuracy. Conversely, other methods aimed at mitigating popularity bias, particularly *MUL*, have typically resulted in more substantial reductions in $nDCG$ scores, similar to observations from the MLM dataset.

In the case of *APLT*, which initially stood at 0.079, the *EquiRate* framework has exhibited remarkable improvements when employing high α values, such as a notable increase to 0.664 when the SRI is configured as HIFA, $\alpha = 0.9$, and SRG is set as $GRV-i_{avg}$. This enhancement suggests a strategic shift of the recommendation algorithm towards promoting items from the long-tail, thereby enriching the recommendations with a broader spectrum of content. Similarly, for the metric of *Novelty*, which had a baseline of 0.715, enhancements have been recorded across various *EquiRate* configurations. In particular,

the aforementioned variant achieved the highest uplift in *Novelty*, reaching up to 0.796, indicating an inclination towards recommending more novel items to users. With regard to *LTC*, initially quantified at 0.016, a substantial rise to 0.658 was observed in the same variant, signifying an expanded representation of *long-tail* content in the recommendations, thus amplifying the exposure of *niche* items. As for *Entropy*, which was originally at 0.457, there's been a general trend of increase with *EquiRate* adaptations, especially with the variant in question peaking at 0.808. This points to a diversification in the recommendation lists, offering users an enriched selection of options. Ultimately, the application of *EquiRate* and its various adaptations proves to be an effective strategy for mitigating popularity bias and fostering diversity within the DB dataset.

In contrast to the MLM dataset, when examining *Entropy* and *LTC*, it is observed that the *EquiRate* variant with the highest performance markedly surpasses other methods aimed at reducing popularity bias, as evidenced by Table 4.7. Conversely, in terms of other metrics (i.e., *nDCG*, *APLT*, and *Novelty*), this leading *EquiRate* variant is capable of achieving results that are on par with those of benchmark methods.

The analysis of improvement rates in the *FusionIndex* illustrated in Fig. 4.23 demonstrates that all the *EquiRate* variants considered surpass the benchmark methods for reducing popularity bias in the DB dataset. Specifically, variants HIFA (HIFA (0.9)_GRV- i_{avg} , HIFA (0.5)_GRV- i_{avg}) and TPA (0.9)_GRV- i_{avg} have achieved the top *FusionIndex* scores, exhibiting outstanding performance improvements of 31.906%, 30.085%, and 29.068%, respectively. These findings confirm that *EquiRate* significantly mitigates popularity bias and markedly increases diversity in the recommendation lists for the DB dataset. Notably, the *LNSF* and *MUL* methods have been observed to substantially reduce the *FusionIndex* of recommendations, primarily due to their significant reduction in overall accuracy, which consequently deteriorates the *FusionIndex*. This pattern observed with *LNSF* is also applicable to the MLM dataset, as depicted in Fig. 4.22.

When examining the accuracy and beyond-accuracy metrics for the Yelp dataset as depicted in Table 4.24, it's observed that for *nDCG*, initially at 0.046, most *EquiRate* variants have generally elevated this metric. This stands in contrast to the ML and DB datasets, where the introduction of *EquiRate* variants resulted in minimal reductions in *nDCG* values, as previously outlined. Considering the trade-off between accuracy and

Table 4.8. The assessment of the *EquiRate* variants in comparison with various popularity-debiasing techniques based on *nDCG*, *APLT*, *Novelty*, *LTC*, and *Entropy* metrics specifically for the Yelp dataset.

<i>Debiasing Method</i>	<i>SRI</i>	α	<i>SRG</i>	<i>Approach</i>	<i>nDCG</i>	<i>APLT</i>	<i>Novelty</i>	<i>LTC</i>	<i>Entropy</i>
<i>Original</i>					0.046	0.072	0.964	0.014	0.255
<i>EquiRate</i>	TPA	0.9	PRV		0.057	0.694	0.964	0.118	0.275
	HIFA	0.5	PRV		0.057	0.709	0.968	0.105	0.266
	TPA	0.5	PRV		0.056	0.443	0.964	0.063	0.255
	TPA	0.5	RVG	<i>mid</i>	0.059	0.360	0.964	0.014	0.265
	HIFA	0.5	RVG	<i>mid</i>	0.056	0.511	0.967	0.021	0.259
	TPA	0.1	GRV	<i>i_{avg}</i>	0.054	0.448	0.970	0.028	0.266
	OPA	0.1	GRV	<i>i_{avg}</i>	0.061	0.156	0.963	0.015	0.262
	HIFA	0.9	RVG	<i>mid</i>	0.053	0.533	0.970	0.034	0.255
	OPA	0.5	GRV	<i>i_{avg}</i>	0.056	0.476	0.966	0.018	0.250
	TPA	0.9	RVG	<i>mid</i>	0.055	0.429	0.965	0.019	0.252
<i>xQuad</i>					0.047	0.496	0.972	0.002	0.286
<i>LNSF</i>					0.051	0.562	0.979	0.001	0.277
<i>PAW</i>					0.055	0.001	0.964	0.001	0.256
<i>AUG</i>					0.028	0.050	0.975	0.001	0.262
<i>MUL</i>					0.010	0.797	0.990	0.001	0.261

beyond-accuracy measures in recommendation systems, the accuracy enhancements achieved through *EquiRate* variants unfortunately do not correspond to significant gains in *Entropy* and *Novelty*, as indicated in Table 4.24. Nonetheless, it can be deduced that all *EquiRate* variants significantly improve the *APLT* for recommendations, and some (e.g., TPA (0.9)_PRV) yield substantial improvements in *LTC*. In summary, *EquiRate* variants effectively increase the probability of incorporating tail items in the final recommendations by addressing popularity bias and delivering more precise recommendations than the original algorithm.

Furthermore, Table 4.24 illustrates that the top-performing *EquiRate* variant significantly surpasses other debiasing methods in terms of the *nDCG* and *LTC* metrics. Conversely, for other metrics (e.g., notably *Novelty* and *Entropy*), the highest achieving *EquiRate* variant delivers outcomes on par with benchmark methods. Additionally, the *FusionIndex*

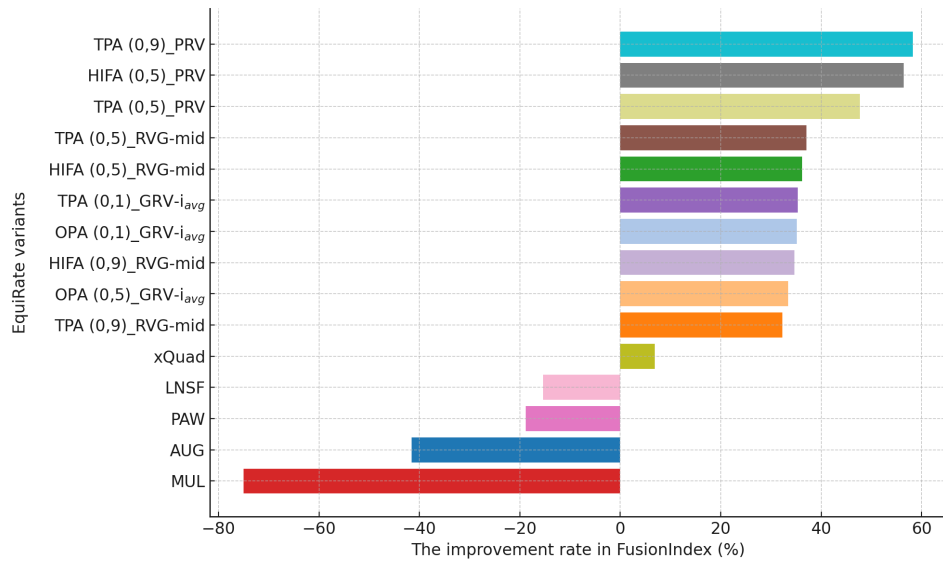


Figure 4.24. The enhancement in the *FusionIndex* scores following the application of popularity-debiasing techniques for the Yelp dataset.

scores, as showcased in Fig. 4.24, reveal that the *EquiRate* approaches, especially those employing the TPA strategy, achieve the topmost *FusionIndex* scores within the Yelp dataset. It's also crucial to highlight that the level of enhancement in the *FusionIndex* for the Yelp dataset markedly exceeds that seen in the MLM and DB datasets. This discovery underscores the applicability of the *EquiRate* in real-life contexts, given that the Yelp dataset's sparser and broader nature presents a scenario closer to real-world conditions.

On the other hand, all benchmark debiasing methods generally achieve lower *FusionIndex* scores in the Yelp dataset compared to the *EquiRate*, as shown in Fig. 4.24. Significantly, except for *xQuad*, they fail to enhance the *FusionIndex* scores, with some methods like *MUL* even resulting in notable declines. Existing popularity-debiasing techniques are predominantly designed to ameliorate certain facets of recommendations by addressing popularity bias. However, evaluations based on proposed *FusionIndex* metric indicate that conventional debiasing approaches might not be as effective when simultaneously addressing accuracy and various well-known beyond-accuracy aspects of recommendations.

5. RESULTS AND DISCUSSION

Based on the comprehensive experiments conducted using three well-known datasets, present the most significant insights obtained as follows.

- Widely adopted privacy-protection methods within recommendation systems can effectively alleviate users' privacy concerns while concurrently achieving a marked reduction in the popularity levels of the items recommended.
- When it comes to generating recommendations that excel in dimensions beyond mere accuracy, such as diversity, novelty, equality, and coverage of the entire catalogue, the algorithms that stand out for their effectiveness are *IBPR* and *WMF*. In contrast, those that are predominantly less effective tend to be the ones inspired by neural networks, notably *NEUMF* and *VAECF*.
- Enhancements in the privacy levels accorded to users correlate with improvements in the quality of their recommendations, specifically in terms of diversity, novelty, coverage of the catalogue, and equity in the representation of items.
- In scenarios where privacy protection is applied to traditional recommendation systems, the *WMF* algorithm stands out as the most effective, balancing both accuracy and beyond-accuracy measures. Moreover, the algorithms *VAECF* and *WBPR* consistently deliver the most precise recommendations.
- With the enhancement of privacy levels for users, there is a noticeable decline in the precision of the recommendations they receive.
- The efficacy of popularity-debiasing strategies is enhanced in generating recommendations with reduced bias towards item popularity as the level of privacy protection for users is elevated. Nonetheless, the impact of these methods on both the accuracy and the quality of recommendations beyond accuracy significantly wanes for users who place a higher emphasis on their privacy. For users with an acute sensitivity to privacy, termed *fiery fundamentalists*, certain debiasing approaches may even become ineffectual.

- Variants of the *EquiRate* that employ HIFA or TPA as the SRI strategy, rather than OPA, demonstrate a greater capacity to diminish popularity bias and facilitate recommendations of superior quality. This is attributable to the fact that HIFA and TPA specifically target head items in the catalog for computing the average popularity score, which subsequently informs the volume of ratings to be injected, as elaborately discussed in Section 4.5. Conversely, the OPA strategy accounts for the average popularity across the entire catalog to determine the injection of ratings. Consequently, employing HIFA and TPA results in a larger infusion of ratings into the profiles of tail items compared to OPA, thereby more effectively countering popularity bias through these methodologies.
- The *EquiRate* methodology is adept at tackling the issue of popularity bias and generating recommendations of a higher caliber, particularly from the perspective of beyond-accuracy metrics, when configured with elevated α values. The heightened efficiency is attributed to the increased α values leading to the injection of a larger volume of synthetic ratings into the profiles of chosen items. Such a strategy is instrumental in mitigating the discrepancies in rating distribution present within the dataset, which are the fundamental contributors to the popularity bias problem encountered in recommendation systems.
- When evaluating SRG strategies, *EquiRate* variations tend to outperform when employing the GRV strategy. This method, which calculates the average ratings of items (denoted as i_{avg}), as detailed in Section 4.5, is more successful than its counterparts. The strength of the GRV strategy resides in its capacity to maintain the inherent rating patterns of item profiles intact, avoiding alterations caused by the insertion of average ratings.
- Research findings indicate that conventional methods for mitigating popularity bias demonstrate inconsistent performances across various beyond-accuracy evaluation metrics, even as they tackle popularity bias. This inconsistency arises because these methods are designed to improve specific facets of recommendation systems, which complicates achieving a holistic assessment. In contrast, the *FusionIndex* metric provides a coherent and balanced framework for assessing both accuracy and beyond-

accuracy characteristics of recommendation systems. The use of the *FusionIndex* in experiments reveals that traditional debiasing techniques may occasionally fall short, as they do not enhance the *FusionIndex* scores of recommendations and might even lead to reductions.

- The top-performing variants of *EquiRate* generally outshine other debiasing techniques, despite sometimes underperforming in specific evaluation metrics. However, they consistently demonstrate substantial improvements in the *FusionIndex* scores of recommendations when compared to traditional strategies. This pattern is observable across all tested datasets, with notable emphasis on the Yelp dataset. It is essential to acknowledge that the Yelp dataset, characterized by its extensive dimensionality and high sparsity, more accurately reflects real-world conditions. This underlines the *EquiRate* method's capacity to significantly enhance the diversity and fairness of recommendations, particularly in practical and complex scenarios.

In summary, comprehensive studies on three widely recognized datasets indicate that the *EquiRate* method, particularly through its HIFA and TPA strategies, significantly mitigates the issue of popularity bias within recommendation systems. These approaches outperform the OPA strategy by effectively leveraging data from popular items to refine recommendation outputs. The emphasis on higher α values plays a pivotal role in improving the overall quality of recommendations, with a special focus on metrics that measure beyond mere accuracy. The GRV strategy is particularly noted for its ability to maintain authentic rating patterns, thereby avoiding the introduction of artificial distortions. Additionally, this research highlights the limitations inherent in traditional debiasing techniques and underscores the comprehensive evaluative power of the *FusionIndex*. Notably, *EquiRate* showcases robust effectiveness across various datasets, with its impact most prominently observed in the Yelp dataset, underlining its applicability in real-world settings and its contribution towards fostering more equitable and diverse recommendation systems.

6. CONCLUDING REMARKS

One of the primary objectives of recommender systems is to offer relevant recommendations to users while safeguarding their privacy. The predominant strategy to achieve this is by obscuring the list of items that have been rated and the actual ratings assigned to these items. Subsequently, collaborative recommendation models are trained on anonymized preference data instead of the genuine user-item interaction matrix. Given that the notable popularity bias issue in recommender systems primarily stems from the imbalances in the distribution of ratings, such privacy-preserving measures could substantially alter the extent of bias detected in the recommendations. More critically, it may render existing solutions for addressing this bias considerably less effective.

In this research, primarily explored the variations in the inclination towards popularity bias within recommender systems for four distinct user personas categorized based on different levels of privacy protection. Additionally, scrutinize how the efficacy of three established popularity-debiasing techniques shifts when considering user personas delineated by privacy levels. A pivotal discovery from extensive experimental studies indicates that concealing individuals' actual preference data not only mitigates their privacy concerns but concurrently reduces the bias towards item popularity in recommendations. In essence, enhancing user privacy makes collaborative recommendation systems more immune to the popularity bias issue, allowing users who are more privacy-conscious to access higher-quality recommendations concerning aspects beyond mere accuracy, such as diversity, novelty, catalog coverage, and fairness in item representation, albeit at a tolerable accuracy trade-off. Another significant observation from this analysis reveals that, although the effectiveness of specific popularity-debiasing methods in promoting less popular items in recommendations improves with heightened user privacy protection, their impact on the quality of recommendations, both in terms of accuracy and beyond-accuracy metrics, significantly wanes. Some methods may even become impractical for users who are extremely sensitive about their privacy.

When developing user personas based on privacy levels, acknowledge randomized perturbation techniques as a prevalent strategy to ensure user privacy. Nonetheless, investigation could be broadened to include alternative techniques for preserving privacy, such

as differential privacy and cryptographic methods. Furthermore, there is a demand for innovative popularity-debiasing approaches that remain unaffected by the data masking process, thereby delivering consistent performance in scenarios where users' privacy is safeguarded.

Popularity bias in recommendation systems tends to disproportionately favor well-known items, often to the detriment of lesser-known or *niche* items. This preference can restrict the diversity of recommendations and hinder the exposure of a wider variety of content, adversely affecting the user experience. Addressing this issue is critical for achieving a more equitable and comprehensive exposure of items, which can enhance user satisfaction by promoting the discovery of previously overlooked content and ensuring a fair platform for all content creators.

The *EquiRate* popularity debiasing technique introduced in this research provides a recalibrated approach to recommendations by injecting synthetic ratings into lesser-known or "tail" items, ensuring equitable representation. This technique enhances the visibility of typically overlooked items, offering them a fair opportunity to be recommended. The *EquiRate* employs the renowned *Pareto principle* for item classification and implements three *Synthetic Rating Injection (SRI)* strategies, namely *Overall Popularity Adjustment (OPA)*, *Head Item Focused Adjustment (HIFA)*, and *Threshold-Based Popularity Adjustment (TPA)*, to strategically allocate and calculate synthetic ratings. Moreover, three *Synthetic Rating Generation (SRG)* strategies, *Generalized Rating Value (GRV)*, *Predictive Rating Value (PRV)*, and *Randomized Value Generation (RVG)*, are utilized to maintain the genuine profile of items while reducing popularity bias. The nine variants of *EquiRate*, derived from the combinations of *SRI* and *SRG* strategies, are subjected to experimental validation to determine the most effective debiasing variant. The chosen variant pre-processes the data for recommendation algorithms, leading to a balanced dataset that improves the representation of less popular items and increases the diversity, novelty, and fairness of the recommendations, fostering a more equitable ecosystem. Additionally, the *FusionIndex* is introduced as an advanced metric for evaluating recommendation lists by simultaneously measuring accuracy and beyond-accuracy aspects such as diversity, catalog coverage, novelty, and fairness. This comprehensive metric is invaluable for assessing the impact of strategies aimed at mitigating popularity bias in recommendation systems.

The experimental results from the well-regarded MovieLens-1M (MLM), Douban-Book (DB), and Yelp datasets offer insightful evaluations for the *EquiRate* approach against established popularity debiasing methods. Within these datasets, various *EquiRate* configurations substantially surpass the performance of five notable popularity debiasing methods regarding the *FusionIndex* measure. Specifically, the *HIFA-GRV* variant of *EquiRate* consistently demonstrates the most significant improvement in the MLM and DB datasets, effectively addressing the issue of popular item overrepresentation. Conversely, the *TPA-PRV* variant shows superior performance in the Yelp dataset, underlining its suitability for larger and sparser datasets. These findings underscore the effectiveness of *EquiRate* in mitigating popularity bias and augmenting recommendation diversity. Furthermore, while certain existing approaches to minimizing popularity bias in recommendation systems exhibit encouraging outcomes on specific metrics beyond accuracy, they generally achieve lower *FusionIndex* scores compared to the best-performing *EquiRate* variants. Notably, when assessed using the *FusionIndex*, these methods often significantly underperform in harmonizing accuracy with metrics beyond accuracy, failing to improve the *FusionIndex* scores of their recommendations and, in some instances, even leading to marked reductions.

Proposed three directions for future research related to this study. The first direction involves investigating alternative algorithms for calculating the *Predictive Rating Value* (*PRV*) within the Synthetic Rating Injection framework, where the VAECF algorithm is currently implemented. The second suggestion aims at enriching and advancing the *FusionIndex* metric by integrating additional measures, such as the average popularity of recommended items. This improvement seeks to broaden the scope of *FusionIndex*, making it a more encompassing and reflective metric for future research efforts. Lastly, leveraging oversampling strategies from traditional machine learning could prove beneficial in determining the value of ratings injected into less popular, or "tail" items.

7. REFERENCES

- Abdollahpouri, H. (2020). Popularity bias in recommendation: A multi-stakeholder perspective (PhD thesis, University of Colorado at Boulder).
- Abdollahpouri, H., & Burke, R. (2019). Multi-stakeholder recommendation and its connection to multi-sided fairness. arXiv preprint arXiv:1907.13158.
- Abdollahpouri, H., Burke, R., & Mobasher, B. (2017). Controlling popularity bias in learning-to-rank recommendation. In Proceedings of the Eleventh ACM Conference on Recommender Systems (pp. 42-46).
- Abdollahpouri, H., Burke, R., & Mobasher, B. (2018). Popularity-aware item weighting for long-tail recommendation. arXiv preprint arXiv:1802.05382.
- Abdollahpouri, H., Burke, R., & Mobasher, B. (2019). Managing popularity bias in recommender systems with personalized re-ranking. arXiv preprint arXiv:1901.07555.
- Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2019). The unfairness of popularity bias in recommendation. arXiv preprint arXiv:1907.13286.
- Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2020). The connection between popularity bias, calibration, and fairness in recommendation. In Fourteenth ACM Conference on Recommender Systems (pp. 726-731).
- Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B., & Malthouse, E. (2021). User-centered evaluation of popularity bias in recommender systems. In Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (pp. 119-129).
- Agarwal, A., Mishra, D. S., & Kolekar, S. V. (2022). Knowledge-based recommendation system using semantic web rules based on learning styles for MOOCs. *Cogent Engineering*, 9(1), 2022568.

- Agrawal, D., & Aggarwal, C. C. (2001). On the design and quantification of privacy preserving data mining algorithms. In ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems.
- Ahanger, A. B., Aalam, S. W., Bhat, M. R., & Assad, A. (2022). Popularity bias in recommender systems-a review. In *Emerging Technologies in Computer Engineering: Cognitive Computing and Intelligent IoT: 5th International Conference, ICETCE 2022, Jaipur, India, February 4–5, 2022, Revised Selected Papers* (pp. 431–444). Springer.
- Antikacioglu, A., & Ravi, R. (2017). Post processing recommender systems for diversity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 707–716).
- Badsha, S., Yi, X., & Khalil, I. (2016). A practical privacy-preserving recommender system. *Data Science and Engineering*, 1(3), 161–177.
- Badsha, S., Yi, X., Khalil, I., & Bertino, E. (2017). Privacy preserving user-based recommender system. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)* (pp. 1074–1083). IEEE.
- Badsha, S., Yi, X., Khalil, I., Liu, D., Nepal, S., & Lam, K.-Y. (2018). Privacy preserving user based web service recommendations. *IEEE Access*, 6, 56647–56657.
- Batmaz, Z., Yurekli, A., Bilge, A., & Kaleli, C. (2019). A review on deep learning for recommender systems: challenges and remedies. *Artificial Intelligence Review*, 52(1), 1–37.
- Bilge, A., & Polat, H. (2013). A scalable privacy-preserving recommendation scheme via bisecting k-means clustering. *Information Processing & Management*, 49(4), 912–927.
- Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-based Systems*, 46, 109–132.

- Boratto, L., Fenu, G., & Marras, M. (2019). The effect of algorithmic bias on recommender systems for massive open online courses. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I* (pp. 457–472). Springer.
- Boratto, L., Fenu, G., & Marras, M. (2021). Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Information Processing & Management*, 58(1), 102387.
- Borges, R., & Stefanidis, K. (2021). On mitigating popularity bias in recommendations via variational autoencoders. In *Proceedings of the 36th annual ACM symposium on applied computing* (pp. 1383–1389).
- Bosri, R., Rahman, M. S., Bhuiyan, M. Z. A., & Omar, A. A. (2020). Integrating blockchain with artificial intelligence for privacy-preserving recommender systems. *IEEE Transactions on Network Science and Engineering*, 8(2), 1009–1018.
- Boutet, A., Frey, D., Guerraoui, R., Jégou, A., & Kermarrec, A.-M. (2016). Privacy-preserving distributed collaborative filtering. *Computing*, 98(8), 827–846.
- Casino, F., Domingo-Ferrer, J., Patsakis, C., Puig, D., & Solanas, A. (2015). A k-anonymous approach to privacy preserving collaborative filtering. *Journal of Computer and System Sciences*, 81(6), 1000–1011.
- Chen, C., Zhang, M., Liu, Y., & Ma, S. (2018). Missing data modeling with user activity and item popularity in recommendation. In *Information Retrieval Technology: 14th Asia Information Retrieval Societies Conference, AIRS 2018, Taipei, Taiwan, November 28-30, 2018, Proceedings 14* (pp. 113–125). Springer.
- Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2020). Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240*.
- Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2023). Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3), 1–39.

- Chen, X., & Huang, V. (2012). Privacy preserving data publishing for recommender system. In 2012 IEEE 36th Annual Computer Software and Applications Conference Workshops (pp. 128–133). IEEE.
- Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9), e2023301118.
- Dong, Q., Xie, S.-S., & Li, W.-J. (2021). User-item matching for recommendation fairness. *IEEE Access*, 9, 130389–130398.
- Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265–284). Springer.
- Elahi, M., Khosh Kholgh, D., Kiarostami, M. S., Saghari, S., Parsa Rad, S., & Tkal-cic, M. (2021). Investigating the impact of recommender systems on user-based and item-based popularity bias. *Information Processing & Management*, 58(5), 102655.
- Elmisery, A. M., & Botvich, D. (2017). An enhanced middleware for collaborative privacy in iptv recommender services. *arXiv preprint arXiv:1711.07593*.
- Gangwar, A., & Jain, S. (2021). An adaptive boosting technique to mitigate popularity bias in recommender system. *arXiv preprint arXiv:2109.05677*.
- Gantner, Z., Drumond, L., Freudenthaler, C., & Schmidt-Thieme, L. (2012). Personalized ranking for non-uniformly sampled items. In *Proceedings of KDD Cup 2011* (pp. 231–247). PMLR.
- Gong, S. (2011). Privacy-preserving collaborative filtering based on randomized perturbation techniques and secure multiparty computation. *International Journal of Advancements in Computing Technology*, 3(4), 89–99.
- Gopalan, P., Hofman, J. M., & Blei, D. M. (2015). Scalable recommendation with hierarchical poisson factorization. In *UAI* (pp. 326–335).

- Guerraoui, R., Kermarrec, A.-M., Patra, R., & Taziki, M. (2015). D 2 p: distance-based differential privacy in recommenders. *Proceedings of the VLDB Endowment*, 8(8), 862–873.
- Gulsoy, M., Yalcin, E., & Bilge, A. (2023). Robustness of privacy-preserving collaborative recommenders against popularity bias problem. *PeerJ Computer Science*, 9, e1438.
- Gupta, S., Kaur, K., & Jain, S. (2023). Eqbal-rs: Mitigating popularity bias in recommender systems. *Journal of Intelligent Information Systems*, 1–26.
- Harper, F. M., & Konstan, J. A. (2015). The movielens datasets: History and context. *ACM transactions on interactive intelligent systems (tiis)*, 5(4), 1–19.
- He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web* (pp. 173–182).
- Hou, L., Pan, X., & Liu, K. (2018). Balancing the popularity bias of object similarities for personalised recommendation. *The European Physical Journal B*, 91, 1–7.
- Hou, M., Wei, R., Wang, T., Cheng, Y., & Qian, B. (2018). Reliable medical recommendation based on privacy-preserving collaborative filtering. *Computers, Materials & Continua*, 56(1), 137–149.
- Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining* (pp. 263–272). IEEE.
- Jannach, D., Kamehkhosh, I., & Bonnin, G. (2016). Biases in automated music playlist generation: A comparison of next-track recommending techniques. In *Proceedings of the 2016 conference on user modeling adaptation and personalization* (pp. 281–285).
- Jannach, D., Lerche, L., Kamehkhosh, I., & Jugovac, M. (2015). What recommenders recommend: An analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction*, 25, 427–491.

- Jannach, D., & Zanker, M. (2022). Value and impact of recommender systems. In *Recommender Systems Handbook* (pp. 519–546). Springer.
- Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2014). Correcting popularity bias by enhancing recommendation neutrality. *RecSys Posters*, 805.
- Kearns, M., & Ron, D. (1997). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Proceedings of the tenth annual conference on Computational learning theory* (pp. 152–162).
- Kerkouche, R., Acs, G., Castelluccia, C., & Genevès, P. (2021). Privacy-preserving and bandwidth-efficient federated learning: An application to in-hospital mortality prediction. In *Proceedings of the Conference on Health, Inference, and Learning* (pp. 25–35).
- Ko, H., Lee, S., Park, Y., & Choi, A. (2022). A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1), 141.
- Koren, Y., Rendle, S., & Bell, R. (2022). Advances in collaborative filtering. In *Recommender Systems Handbook* (pp. 91–142). Springer.
- Kowald, D., Schedl, M., & Lex, E. (2020). The unfairness of popularity bias in music recommendation: A reproducibility study. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42* (pp. 35–42). Springer.
- Le, D. D., & Lauw, H. W. (2017). Indexable bayesian personalized ranking for efficient top-k recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 138).
- Lesota, O., Melchiorre, A., Rekabsaz, N., Brandl, S., Kowald, D., Lex, E., & Schedl, M. (2021). Analyzing item popularity bias of music recommender systems: are different genders equally affected? In *Proceedings of the 15th ACM Conference on Recommender Systems* (pp. 601–606).

- Li, D., Chen, C., Lv, Q., Shang, L., Zhao, Y., Lu, T., & Gu, N. (2016). An algorithm for efficient privacy-preserving item-based collaborative filtering. *Future Generation Computer Systems*, 55, 311–320.
- Li, D., Lv, Q., Shang, L., & Gu, N. (2017). Efficient privacy-preserving content recommendation for online social communities. *Neurocomputing*, 219, 440–454.
- Li, L., Chen, J., & Raghunathan, S. (2018). Recommender system rethink: Implications for an electronic marketplace with competing manufacturers. *Information Systems Research*, 29(4), 1003–1023.
- Liang, D., Krishnan, R. G., Hoffman, M. D., & Jebara, T. (2018). Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference* (pp. 689–698).
- Lin, L., Tian, Y., & Liu, Y. (2021). A blockchain-based privacy-preserving recommendation mechanism. In *2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP)* (pp. 74–78). IEEE.
- Liu, X., Liu, A., Zhang, X., Li, Z., Liu, G., Zhao, L., & Zhou, X. (2017). When differential privacy meets randomized perturbation: a hybrid approach for privacy-preserving recommender system. In *International Conference on database systems for advanced applications* (pp. 576–591). Springer.
- Liu, Z., Mei, S., Xiong, C., Li, X., Yu, S., Liu, Z., Gu, Y., & Yu, G. (2023). Text matching improves sequential recommendation by reducing popularity biases. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (pp. 1534–1544).
- Lu, S., Gao, Z., Xu, Q., Jiang, C., Zhang, A., & Wang, X. (2022). Class-imbalance privacy-preserving federated learning for decentralized fault diagnosis with biometric authentication. *IEEE Transactions on Industrial Informatics*, 18(12), 9101–9111.
- Ma, Z., & Dong, Q. (2021). Alleviating the unfairness of recommendation by eliminating

- the conformity bias. In *Journal of Physics: Conference Series*, volume 2004, page 012008. IOP Publishing.
- McSherry, F., & Mironov, I. (2009). Differentially private recommender systems: Building privacy into the netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 627–636).
- Naghiaei, M., Rahmani, H. A., & Dehghan, M. (2022). The unfairness of popularity bias in book recommendation. In *Advances in Bias and Fairness in Information Retrieval: Third International Workshop, BIAS 2022, Stavanger, Norway, April 10, 2022, Revised Selected Papers* (pp. 69–81). Springer.
- Omar, A. A., Bosri, R., Rahman, M. S., Begum, N., & Bhuiyan, M. Z. A. (2019). Towards privacy-preserving recommender system with blockchains. In *Dependability in Sensor, Cloud, and Big Data Systems and Applications: 5th International Conference, DependSys 2019, Guangzhou, China, November 12–15, 2019, Proceedings 5* (pp. 106–118). Springer.
- Ozturk, A., & Polat, H. (2015). From existing trends to future trends in privacy-preserving collaborative filtering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(6), 276–291.
- Parameswaran, R., & Blough, D. M. (2007). Privacy preserving collaborative filtering using data obfuscation. In *2007 IEEE International Conference on Granular Computing (GRC 2007)* (pp. 380–380). IEEE.
- Park, Y.-J., & Tuzhilin, A. (2008). The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems* (pp. 11–18).
- Polat, H., & Du, W. (2003). Privacy-preserving collaborative filtering using randomized perturbation techniques. In *Third IEEE International Conference on Data Mining* (pp. 625–628). IEEE.

- Polat, H., & Du, W. (2005). Privacy-preserving collaborative filtering. *International journal of electronic commerce*, 9(4), 9–35.
- Polat, H., & Du, W. (2005). Privacy-preserving collaborative filtering on vertically partitioned data. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 651–658). Springer.
- Polatidis, N., Georgiadis, C. K., Pimenidis, E., & Mouratidis, H. (2017). Privacy-preserving collaborative recommendations based on random perturbations. *Expert Systems with Applications*, 71, 18–25.
- Pramod, D. (2022). Privacy-preserving techniques in recommender systems: state-of-the-art review and future research agenda. *Data Technologies and Applications* (ahead-of-print).
- Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: introduction and challenges. In *Recommender Systems Handbook* (pp. 1–34). Springer.
- Salah, A., Rogovschi, N., & Nadif, M. (2016). A dynamic collaborative filtering system via a weighted clustering approach. *Neurocomputing*, 175, 206–215.
- Salah, A., Truong, Q.-T., & Lauw, H. W. (2020). Cornac: A comparative framework for multimodal recommender systems. *J. Mach. Learn. Res.*, 21(95), 1–5.
- Sánchez, P. (2019). Exploiting contextual information for recommender systems oriented to tourism. In *Proceedings of the 13th ACM conference on recommender systems* (pp. 601–605).
- Sanders, R. (1987). The pareto principle: its use and abuse. *Journal of Services Marketing*, 1(2), 37–40.
- Shambour, Q., Hussein, A. A., Kharma, Q., & Abu-Alhaj, M. M. (2022). Effective hybrid content-based collaborative filtering approach for requirements engineering. *Comput. Syst. Sci. Eng.*, 40(1), 113–125.

- Shen, Y., & Jin, H. (2014). Privacy-preserving personalized recommendation: An instance-based approach via differential privacy. In 2014 IEEE International Conference on Data Mining (pp. 540–549). IEEE.
- Shi, C., Hu, B., Zhao, W. X., & Philip, S. Y. (2018). Heterogeneous information network embedding for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 31(2), 357–370.
- Shmueli, E., & Tassa, T. (2017). Secure multi-party protocols for item-based collaborative filtering. In *Proceedings of the eleventh ACM conference on recommender systems* (pp. 89–97).
- Si, M., & Li, Q. (2020). Shilling attacks against collaborative recommender systems: a review. *Artificial Intelligence Review*, 53(1), 291–319.
- Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009, 2009.
- Su, X., Li, P., & Zhu, X. (2023). The influence of herd mentality on rating bias and popularity bias: A bi-process debiasing recommendation model based on matrix factorization. *Behavioral Sciences*, 13(1), 63.
- Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., & Zhou, Y. (2019). A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM workshop on artificial intelligence and security* (pp. 1–11).
- Wang, Y., Ma, W., Zhang, M., Liu, Y., & Ma, S. (2023). A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3), 1–43.
- Wei, R., Tian, H., & Shen, H. (2018). Improving k-anonymity based privacy preservation for collaborative filtering. *Computers & Electrical Engineering*, 67, 509–519.
- Weimer, M., Karatzoglou, A., & Smola, A. (2008). Improving maximum margin matrix factorization. *Machine Learning*, 72(3), 263–276.

- Xu, R., Baracaldo, N., Zhou, Y., Anwar, A., & Ludwig, H. (2019). Hybridalpha: An efficient approach for privacy-preserving federated learning. In *Proceedings of the 12th ACM workshop on artificial intelligence and security* (pp. 13–23).
- Yalcin, E. (2022). Exploring potential biases towards blockbuster items in ranking-based recommendations. *Data Mining and Knowledge Discovery*, 36(6), 2033–2073.
- Yalcin, E. (2022). Pophybrid: a novel item popularity-aware hybrid approach for long-tail recommendation. In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1–6). IEEE.
- Yalcin, E., & Bilge, A. (2021). Investigating and counteracting popularity bias in group recommendations. *Information Processing & Management*, 58(5), 102608.
- Yalcin, E., & Bilge, A. (2022). Evaluating unfairness of popularity bias in recommender systems: A comprehensive user-centric analysis. *Information Processing & Management*, 59(6), 103100.
- Yalcin, E., & Bilge, A. (2023). Popularity bias in personality perspective: An analysis of how personality traits expose individuals to the unfair recommendation. *Concurrency and Computation: Practice and Experience*, 35, e7647.
- Yargic, A., & Bilge, A. (2019). Privacy-preserving multi-criteria collaborative filtering. *Information Processing & Management*, 56(3), 994–1009.
- Yin, X., Zhu, Y., & Hu, J. (2021). A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)*, 54(6), 1–36.
- Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017). Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 1569–1578).
- Zhang, F., Lee, V. E., & Choo, K.-K. R. (2018). Jo-dpmf: Differentially private matrix factorization learning through joint optimization. *Information Sciences*, 467, 271–281.

- Zhang, Y., Feng, F., He, X., Wei, T., Song, C., Ling, G., & Zhang, Y. (2021). Causal intervention for leveraging popularity bias in recommendation. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 11–20).
- Zou, J., & Fekri, F. (2015). A belief propagation approach to privacy-preserving item-based collaborative filtering. *IEEE Journal of Selected Topics in Signal Processing*, 9(7), 1306–1318.



CURRICULUM VITAE

EDUCATION

Doctorate 2019 – 2024	Akdeniz University, Graduate School of Natural and Applied Sciences, Department of Computer Engineering, Antalya
Master Degree 2018 – 2019	Haliç University, Institute of Graduate Studies, Computer Engineering, Istanbul
Bachelor's Degree 2002 – 2006	Haliç University, Faculty of Engineering, Computer Engineering (Scholarship), Istanbul
Associate Degree 2000 – 2002	Kocaeli University, Köseköy Vocational School, Department of Computer Programming, Kocaeli

PROFESSIONAL AND ADMINISTRATIVE DUTIES

Manager 2020 – Present	Alanya Alaaddin Keykubat University, Distance Education Application and Research Center, Antalya
Project Manager 2020 - 2020	Alanya Alaaddin Keykubat University, BAP Coordination Office, Antalya
Manager 2016 - 2018	Etiya Software Services and Consulting Inc., Service Excellence Directorate, Istanbul
Manager 2015 - 2016	TRT World News Channel, Information Technologies Directorate, Istanbul
Manager 2013 – 2015	Turk Telekom Inc., Information Technologies Directorate, Istanbul
Expert 2012 – 2013	TTNET Inc., Information Technologies Directorate, Istanbul
Engineer 2009 – 2012	Turkcell Communication Services Inc., Application Operation Directorate, Istanbul
Engineer 2007 - 2009	Aksis Computer Services and Consulting Inc., Software Directorate, Istanbul

PUBLICATIONS

Articles published in international peer-reviewed journals

1- Gulsoy, M., Yalcin, E., Bilge, A. (2023). Robustness of privacy-preserving collaborative recommenders against popularity bias problem. PeerJ Computer Science, 9, e1438.

