UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

SENSITIVE ATTRIBUTE ASSOCIATION BIAS IN LATENT FACTOR RECOMMENDATION ALGORITHMS: THEORY AND IN PRACTICE

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

BY

ALEXANDRA BEATTIE
Norman, Oklahoma
2023

SENSITIVE ATTRIBUTE ASSOCIATION BIAS IN LATENT FACTOR RECOMMENDATION ALGORITHMS: THEORY AND IN PRACTICE

A DISSERTATION APPROVED FOR THE
GALLOGLY COLLEGE OF ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Dean Frederick Hougen

Dr. Charles Nicholson

Dr. Talayeh Razzaghi

Dr. Dimitrios Diochnos

Dr. Alexandra Harris-Watson

Acknowledgements

Completing this dissertation marks a significant milestone in my life. I owe immense gratitude to my academic, professional, and personal community for their generous support that has brought me to this point in my journey.

I would like to thank my mentors, advisors, and managers who have supported me along the way. I extend heartfelt appreciation to my doctoral advisor, Dr. Dean Hougen. His belief in my work, invaluable guidance, and mentorship over the past four years have been instrumental in my academic and professional growth. I also want to express my sincere gratitude to my exceptional industrial advisor, Dr. Henriette Cramer. Her dedication and time devoted to nurturing my development as a researcher have been invaluable. Your guidance will forever shape my career in Responsible AI. A special mention goes to my current and past managers at Spotify, Dr. Dan Taber and Josh Baer. Their support, understanding, and constant care for providing research opportunities were instrumental to completing my dissertation and degree. I'm incredibly fortunate to have received such unwavering support and encouragement from such a brilliant group of leaders within the ML space.

I also extend my deepest thanks to my friends and family for their overwhelming support and patience while I worked towards completing my dissertation and earning my degree. Special shout-out to my Dad, I will forever be thankful for

you recognizing my talent and pushing me to the finish line: diamonds are made from heat and pressure! To my fiancé, Reinaldo, thank you for patiently listening to my crying, whining, and threats of quitting during the final days of writing and defending my dissertation. I promise to provide the same level of support as you begin and finish your Ph.D. journey.

Finally, to my brother Adam, you are forever my little dude. Your dedication to serving our country and making the world a safer place for future generations is beyond brave and inspiring. I dedicate this dissertation to you. I am so proud to be your sister, and I cannot believe we are both graduating this academic year. I hope Oklahoma is as good to you as it has been to me. First to fire!

Abstract

This dissertation presents methods for evaluating and mitigating a relatively unexplored bias topic in recommendation systems, which we refer to as *attribute association bias*. Attribute association bias (AAB) can be introduced when leveraging latent factor recommendation models due to their ability to entangle model and implicit attributes into the trained latent space. This type of bias occurs when entity embeddings showcase significant levels of association with specific types of explicit or implicit entity attributes, thus having the potential to introduce representative harms for both consumer and provider stakeholders. We present a novel analysis method framework to help practitioners evaluate their latent factor recommendation models for AAB. This framework consists of three main techniques for gaining insight into sensitive AAB in the recommendation latent space: bias direction creation, bias evaluation metrics, and multi-group evaluation. Methods within our evaluation framework were inspired by techniques presented by the natural language processing research community for measuring gender bias in learned language representations. Additionally, we explore how this bias can be reinforced and produce feedback loops via retraining. Finally, we explore possible mitigation techniques for addressing said bias. Primarily, we demonstrate our methodology with two case studies that evaluate user gender association bias in latent factor recommendation. With our methods, we uncover

the existence of user gender association bias and compare the various methods we propose to help guide practitioners in how best to use our techniques for their systems. In addition to exploring user gender, we experiment with measuring user age association bias as a means for evaluating non-binary AAB.

Contents

List of Figures

List of Tables

Part I

Fairness & Bias in RecSys: An Overview

Chapter 1

Introduction

Recommendation systems are components of algorithms used to serve information or items to downstream users. These recommendation algorithms are tasked with a goal of predicting items based on potential optimization goals, such as relevance, engagement, or revenue. These goals are defined by stakeholders of the recommendation system, which primarily consist of consumers, providers, and side-stakeholders. Recommendation systems often consist of information retrieval and ranking components to create a final product capable of creating a personalized experience for users of the system, as well as delivering value to other system stakeholders.

Over time, recommendation systems have proliferated digital platforms people use every day. Recommendation systems have increased in popularity due to their ability to leverage algorithms to personalize how information is retrieved for a user. Unfortunately, the ability to personalize information access can induce harm to stakeholders of the system [23]. These recommendation-based harms are produced when an algorithm overly relies on sensitive stakeholder features to provide content, items, or information recommendations[23]. Recommendation-

related harms can present themselves in various ways, such as feedback loops, stereotyping by sensitive features (such as gender or race), and fair allocation of information [55, 15, 39].

Fairness in recommendation systems is often referred to as multi-stakeholder fairness due to the various types of people that interact with or are affected by a recommendation system [22]. Multi-stakeholder fairness consists of three main types of fairness: consumer fairness, provider fairness and side-stakeholder fairness [2]. For this dissertation, the focus will be on on provider- and consumer-fairness-related harms in recommendation systems, specifically those induced in latent factor recommendation systems.

In the past, many of these recommendation-related harms have been labeled under the broad category of fair recommendation or, as referenced above, multi-stakeholder recommendation fairness. For the case of this dissertation, we move away from leveraging the term "fair" for evaluation in favor of evaluating "harms". The interpretation of fairness is a highly individual and context-specific undertaking. Given that fairness is a socio-technical term, it is next to impossible to achieve a globally fair system. Fairness is only "achieved" in terms of practitioner-defined constraints on a system. However, researchers and practitioners continue to share so-called "fair" techniques. Even though these methodologies do not guarantee fairness, they do have the potential to reduce harm in relation to fairness concerns. Based on this thought, we present our work in terms of reducing fairness-related harms, not achieving optimal fairness. There have been arguments against focusing on "fairness" due to the idea that achieving true fairness is impossible, thus making the pursuit foolish. We hope that by changing the perspective to lowering fairness-related harms over achieving impossible quantitative fairness, we are able to reframe the problem and highlight

that this area of research wishes to reduce harm to stakeholders.

Latent factor recommendation (LFR) algorithms have become fundamental to industry recommendation settings [21, 8]. These recommendation algorithms, such as collaborative filtering and deep learning, provide predictions of engagement and embedded vector representations of users and items. The resulting trained vector representations can capture entity relationships and characteristics in a condensed dimensional space and allow for comparisons between different entity vectors in the trained latent semantic space.

It has been demonstrated that user and item attributes can become entangled when leveraging these algorithms, resulting in feature duplication and bias amplification [97]. This algorithmic outcome can result in lower and less robust recommendation quality [97]. Research seeking to reduce this type of attribute disentanglement has become increasingly prevalent and showcases favorable results when targeting exposure bias attributed to item attributes, popularity, or user behavior [97, 100, 64, 29, 99, 87]. However, the research primarily focuses on intrinsic mitigation techniques and increasing recommendation performance but does not always provide evaluation techniques to understand how the bias may be captured explicitly or implicitly within the latent space. Attribute disentanglement traditionally requires attributes to be independent and explicitly used in order to implement disentanglement methods. This common requirement results in disentanglement evaluation methods failing to address situations where attributes show interdependence with one another or present themselves implicitly in results. This stipulation hinders processes for identifying systematic bias, such as gender or racial bias, that can be interdependent with other attributes or be implicitly captured by behavior in the recommendation scenario.

Other research concerning this type of bias in representation learning has

shown that systematic bias can occur due to the nature of latent factor algorithms, as previously found in research exploring systematic gender bias in natural language processing (NLP) [41]. NLP research has demonstrated that implicit or systematic bias found in language embeddings can result in downstream representation harms (e.g., translation systems being more likely to generate masculine pronouns when referring to stereotypically-male occupations [75]). NLP researchers have studied this type of bias by evaluating and mitigating *gender association bias* [20, 12, 24]. Even though association bias evaluation has been a focus in other areas of representation learning, it remains largely unstudied for recommendation systems [32].

This dissertation presents a framework for evaluating interdependent and/or systematic bias between recommendation entities. Our framework closes this research gap by evaluating *attribute association bias* resulting from LFR algorithms. AAB is present when entity embeddings showcase significant levels of association with specific types of explicit or implicit entity attributes. For example, while users can be explicitly labeled by gender, pieces of content cannot be gendered. However, due to the potential for attribute entanglement, pieces of content can show measurable levels of implicit association with a gender attribute. Our framework is designed to be attribute agnostic, thus the name attribute association bias (AAB). We showcase this by demonstrating evaluations of user gender and age bias. Our focus on user gender bias examines AAB from a binary approach, while that for user age bias leverages a non-binary perspective. Leveraging the framework, we demonstrate that this risk can also occur within the outputs of latent factor recommendation.

Given the popularity of LFR models in industry systems and the use of their outputs as downstream features in hybrid or multi-component recommendation

systems, it is paramount that practitioners understand and can evaluate AAB to reduce the risk of introducing or reinforcing representation bias in their recommendation systems [8]. Additionally, many of these industry systems leverage hybrid recommendation systems which leverage outputs from previous system components to further fine tune recommendations for consumption [21]. In the case of hybrid LFR systems, embedding outputs are used as features in downstream models. The outputs can also be used in other unrelated modeling scenarios, such as content moderation or ad targeting. If sensitive attribute bias is encoded into the vector, it plausibly can be repeated and amplified when said vectors are used as features in other models. Ignoring this type of bias puts practitioners at risk of unknowingly amplifying stereotypes and representative harm within their recommendation systems. For example, [8] described how matrix factorization algorithms could be combined with "traditional neighborhood-based approaches" to create a recommendation system for Netflix. This combination consisted of LFR embeddings ranked based on neighborhood-based algorithms to produce final recommendations for consumption [8]. If certain sets of user or item vectors were closely associated with a sensitive attribute, the resulting AAB could affect final rankings in the KNN outputs due to groups of vectors forming stereotyped clusters due to this association bias.

Adversarial learning for fair representations has been proposed as a way to reduce how representations created by latent factor recommendation algorithms capture sensitive attributes. However, in previous research, evaluation of reducing this phenomenon primarily leverages standard accuracy or distribution metrics compared between attribute groups. These types of evaluations help showcase how mitigation affects recommendation outcomes but fails to relay how the latent space changes according to the targeted attribute. The AAB metrics

in our framework attempt to fill this gap in research by providing practitioners with the ability to measure how items and users relate to defined sensitive attributes within the trained latent recommendation space. These metrics and evaluation methods can serve as an additional tool for researchers wishing to create "fair" representations beyond the ability for adversarial models to predict a user's sensitive attribute. This one-sided identification of representation bias fails to account for the multi-sided nature of recommendation systems where users and items are often inter-related in entity vector outputs. Understanding AAB is key to not only increasing robustness to privacy-related attacks, but reducing how latent factor recommendation algorithms may capture societal bias in its latent outputs.

Our proposed AAB evaluation framework and mitigation is presented within a novel system-level frameworks we designed, named "SIIM", to showcase the entire process of tackling the complex task of auditing harms and bias in industrial systems [13]. "SIIM", provides essential structure to approaching the ambiguous problem space of auditing for harms and bias in industrial systems [13]. This algorithmic auditing framework can be used to explore other types of bias beyond AAB. The framework can be seen as a "disaggregated evaluation" framework where the focus lies on analyzing AI outputs for harm or bias [70, 13, 11]. This framework consists of four steps: (S)cope, (I)dentify, (I)mplement, and (M)onitor and flag. The first step, scope, addresses the problem of determining "what" to analyze. For example, what sensitive attribute should be analyzed in our LFR model vector output? The second step, identify, focuses on determining the best-suited methodologies for said analysis based on the scope and outputs of the system. The third step, implement, represents the time dedicated to conducting the analysis and determining how to manipulate the data to leverage identified

methods. Finally, the fourth step, monitor and flag, answers the vital question of if significant levels of bias exist within the system. Our dissertation addresses AAB for the majority of these steps, except for providing practical guidance for setting baselines is out of the scope of this paper due to the task's highly context-specific nature.

To the best of our knowledge, we present one of the first evaluation frameworks for addressing AAB (as a type of representation bias) with both an industry and public data case study. Our work provides a practical guide for evaluating AAB in trained vector embeddings. We introduce recommendation entity-specific attribute association vector directions, bias metrics, and evaluation techniques inspired by gender association bias NLP research. Our methods account for differences between recommendation system and NLP representation embeddings and are designed to provide flexibility for evaluation of binary attributes beyond gender bias. Our framework is model and attribute agnostic concerning the type of LFR algorithm. The methods presented can be used for both binary and non-binary attribute settings, but shines particularly for binary comparisons.

This dissertation explores AAB by by introducing, implementing and critiquing our proposed evaluation methods. We also showcase how this type of bias can become reinforced by observing downstream results of classification models and changes in bias in simulated repeated recommendation training. Additionally, we implement various bias mitigation techniques to understand how AAB can be addressed once it has been flagged for mitigation. By exploring both the evaluation and mitigation stages of addressing bias, we provide greater transparency into measuring and mitigating bias in practice for recommendation representations. Evaluation frameworks, such as the one presented in our paper, are essential to practitioners, allowing them to thoroughly investigate the level of bias

in their system before experimenting with and completing expensive mitigation techniques [74]. Even though we do not present novel mitigation methods, our exploration of the process also provides guidance for types of mitigation that may work given a practitioners particular scenario. It is important to note that this dissertation is not an exhaustive exploration of the evaluation and mitigation of AAB, it merely serves as a novel introduction to this type of bias from both an academic and practical point of view.

This dissertations contributions include:

- SIIM as a disaggregated evaluation framework for evaluation bias in practice for recommendation systems
- Definition of AAB within the context of latent factor recommendation algorithms.
- Evaluation methodologies and metrics for analyzing AAB between recommendation entity embeddings.
- Techniques for exploring reinforcing AAB in downstream and subsequent models.
- Exploration of current mitigation methods for addressing AAB.
- Discussion of limitations of this approach and future directions.

Throughout the dissertation, we leverage language such as stereotypes, bias, and harm. When referring to bias, we are discussing algorithmic or qualitative statistically skewed results found in experimental or evaluation settings which can produce harm [84]. We refer to stereotypes as a "product of biases" often held at the societal level, which may or may not be supported in experimental settings

[7]. Stereotyping algorithmic harm occurs when it is produced by algorithmic bias [84]. This type of harm can be seen as "representative" harm due to it reinforcing "the subordination of some groups along the lines of identity" [84]. Our framework, and presented case study, looks to quantitatively evaluate algorithmic bias (AAB) which signals potential for reinforcing stereotypes thus resulting in downstream representative harm. Our evaluation and mitigation of AAB does not target fairness specifically. Instead, it targets how users and items are represented within the latent space, which could cause downstream fairness-related harms. We observe bias as groups of items or users which experience significantly higher levels of related-ness with a specific sensitive attribute as defined by some group of entities.

Chapter 2

Background

As previously described, this dissertation covers a wide breadth of subjects relating to fairness and bias in recommendation systems. In the following sections, we will briefly introduce concepts that are essential building blocks for the research presented in this dissertation. First, we will introduce the concept of recommender systems and describe research areas concerning general recommendations, particularly evaluation in recommendations. Second, we will summarize research concerning fairness and bias in recommendation specific settings. Finally we will cover the concepts that inspired our research, the evaluation of gender association bias in NLP research.

2.1 Recommender Systems

A recommender system has been defined as "software tools and techniques that provide suggestions for items that are most likely of interest to a particular user" [73]. These systems are designed to help users make decisions, like what music to play, clothing to buy, or posts to read [73]. In order to make these decisions,

the system provides personalized recommendations to the user, normally in the form of a ranked list of items [73]. The most basic form of a recommendation system leverages three types of data to create these predictions: users, items and interactions between the users and items [73]. Users are defined as the downstream stakeholders of the system who directly interact with the system predictions to make decisions. Items are the objects that are recommended for the users. The term "item" can refer to anything that is recommended, such as podcasts in a music app, drivers in a rideshare app, or physical items for purchase. The interactions leveraged to train predictions are defined based on the goal of the recommendation system or what is available for training. Interactions can be defined as either explicit or implicit feedback from the user with the system. Explicit feedback most often refers to item ratings given by the user, the three most popular being numerical, oridinal or binary ratings [73]. Implicit feedback is not given directly by the user but is inferred based on how the user interacts with the recommendation system. This type of feedback is often defined by user-item interactions like clicks, shares, purchases, or other user actions which can help the system implicitly understand the users preferences for specific items.

In this section, we will provide a short overview of research areas specific to recommendation systems which are essential to the dissertation. We will focus on commonly used algorithmic techniques for implementing recommendation systems. Next, we will introduce latent factor and hybrid recommendation. Latent factor recommendation can be one (or combinations of) these building-block algorithms to create user- and item-embeddings. Hybrid recommendation also builds upon these algorithms but in the form of creating recommendation components of a recommendation system.

2.1.1 Algorithms

There are a variety of algorithms available for producing recommendations. In this section, we will cover the four main categories of recommendation algorithms found in current research literature.

Collaborative filtering

Collaborative filtering is a technique that leverages similarities between users and content to create recommendations [73]. This technique relies on matrix factorization to learn latent embeddings of the users and items and predict engagement between the two. The algorithm learns to approximate the product of the user embedding matrix and item embedding matrix to the given feedback matrix [46]. Minimizing the objective function reflecting the predicted matrix and true feedback matrix is often done via stochastic gradient descent, or weighted alternating least squares [46]. Collaborative filtering has been called "the most mature and the most commonly implemented" technique for recommendations [73]. However, it has a variety of limitations, such as problems with scaling and providing recommendations for cold-start items and users.

Deep Neural Networks

Deep Neural Network recommendation are defined by their ability to learn deep representations of the content and users in the recommendation environment [96]. The Recommendation System Handbook considers a recommendation algorithm to be a deep neural network if it leverages a "neural differentiable architecture" which "optimizes a differentiable objective function using a variant of stochastic gradient descent" [73]. Types of DNN recommendation algorithms include: mul-

tilayer perceptron, autoencoder, convolutional neural network, recurrent neural network, restricted boltzmann machine, neural autoregressive distribution estimation, adversarial networks, attentional models, graph network models, and deep reinforcement learning [96].

Context aware

Recommender systems are considered context-aware when attributes about the contextual situation are used to produce final recommendations for the user [6]. Context can either be "representational" or "interactional." Representational context is defined by attributes known before the interaction or recommendation occurs and is static [6]. Interactional context is dynamic and can change over time depending on the activity of the user [6]. This context can be injected into the modeling process either pre- or post-filtering [6]. These context-aware recommender systems commonly leverage multiple algorithms to achieve this goal [6].

Content based

Content-based filtering techniques leverage the features of the items to create recommendations based on what the user explicitly likes [61]. These item features are domain-specific and must be engineered for the recommendation task at hand, requiring a fair amount of domain knowledge to label the data correctly [61]. Unlike collaborative filtering, this technique does not look to similar users to create predictions [61]. Instead, it leverages the similarity between items to create the predictions by recommending items most similar to items with which a user positively interacts [61]. Popular algorithms used to learn the relationships between items for content-based filtering include vector space models, probabilistic

models, decision trees, and neural networks [61].

2.1.2 Latent Factor Recommendation

Latent factor recommendation (LFR) algorithms have become fundamental to industry recommendation settings [21, 8]. These recommendation algorithms, such as collaborative filtering and deep learning, provide predictions of engagement and embedded vector representations of users and items. The resulting trained vector representations can capture entity relationships and characteristics in a condensed dimensional space and allow for comparisons between different entity vectors in the trained latent semantic space.

Latent Factor Recommendation algorithms leverage user, item, and interaction data to output entity vectors which can be related to one another in some n-dimensional space. One can assume that any model leveraged for recommendations which maps users and items into the same latent space for final predictions is a latent factor recommendation algorithm. There are a variety of available algorithms for this modeling scenario, but we choose to focus on three seminal deep algorithms leveraged for latent factor recommendation: Bayesian Personalized Ranking (BPR), Neural Collaborative Filtering (NCF), and Deep Matric Factorization (DMF). We chose to focus on deep latent factor recommendation due to its common occurence in industry settings, as showcased with our context-aware deep LFR industry case study and other industry publications [8].

Deep neural (or just neural) networks have become increasingly popular for latent factor recommendation due to their ability to substitute the inner product, leveraged in CF and MF, with a neural architecture. The implementation results in two sets of embeddings for each entity type (users and items). For our analysis,

we evaluate both types of vectors, the general matrix factorization model and the multi layer perceptron model vectors. This differs from DMF, or deep MF which solely leverages one latent vector for an entity.

2.1.3 Hybrid Recommendation

Hybrid recommendation is not necessarily one specific algorithm, but a methodology for building recommendation systems. Hybrid recommenders consist of multiple recommendation components to provide final recommendations to serve to a downstream stakeholder. Burke et al. defines hybrid recommender systems as a modeling system which leverages two or more of the algorithmic methodology groups described above to improve upon final recommendation performance. There are a variety of ways that one can implement a hybrid recommender system, in fact, Burke et al. identifies seven different types. In this dissertation, our exploration focusing on the effects of AAB is most relevant to feature augmentation and meta-level hybrid recommender systems. Feature augmentation hybrid recommendation systems involve one recommender component creating features, most popularly entity embeddings, as an input to a downstream recommendation component [21]. A meta-level hybrid recommender consists of "production line"-esque set of multiple model components with predictions being leveraged as inputs into the subsequent component [21].

These two categories of hybrid recommendation are quite prevalent in industry systems for producing recommendations at scale. Recommender systems in industry settings are commonly designed as hybrid recommendation systems consisting of multiple components to create final predictions [21]. Meta-level hybrid recommendation in practice commonly consists of candidate generation and

ranking components [8]. The candidate generation component acts as a filter of content, creating a final pool of items to be ranked by the ranking or scoring component [8]. The candidate generation component leverages a latent factor recommendation algorithm (such as DNN or collaborative filtering) to create user and item embeddings; these embeddings are used by downstream components to further refine the candidate pool via k-nearest neighbors or providing final rankings with rank-specific algorithms [8].

Recommendation Components

Hybrid recommendation components work together to provide the final recommendations seen by the consumer. These components include generating and retrieving pools of content., filtering said pools by ranking for high-priority goals and re-ranking the final lists of content for consumption. It is essential to make this distinction because different types of components have unique goals for their outputs. Due to the uniqueness of their goals, they may leverage different algorithms, which in turn require unique evaluation metrics and mitigation techniques [53, 79, 93, 78].

Academic literature commonly categorizes the goals of these components into two categories: ranking and rating prediction. These are most often evaluated regarding relevancy (or accuracy) [81]. Rating prediction focuses on "predicting the rating value that a user would assign to an item which s/he has not rated yet" [81]. On the other hand, ranking "is a useful approach when the recommendation task is, for each user, to pick a small number, say N, of items from among all available items in the collection" [81]. Regarding system components, content pool generation often focuses on rating prediction, while ranking and re-ranking focus on picking n number of items for final consumption by the user.

2.2 Fairness & Bias in Recommendation

The evaluation of recommendation systems is an integral area pertaining to the recommender research space. Present research defines a variety of evaluation dimensions for analyzing ones recommendation system or model. The three main dimensions of evaluating a recommendation system are relevance, diversity, and novelty [76]. Relevance measures if the recommendations provide a beneficial utility to the end consumer of the prediction [76]. For example, a prediction in a music recommendation system would be considered relevant if it provided utility to the final consumer. Measuring diversity focuses on evaluating whether the recommendation results consist of multiple item types [76]. In the case of music recommendation, one may measure the diversity of genres recommended to a user in their playlists. Finally, novelty measures the ability of a RecSys to provide a level of serendipity in predictions by serving unexpected but still beneficial results to the consumer [76].

More recently, the evaluation of "fairness" has become it's own evaluation dimension [33]. RecSys are often evaluated in terms of multi-stakeholder fairness. Robin Burke introduced this idea with two definitions: C-fairness and P-fairness [23]. C-fairness refers to the idea of "consumer" fairness, which evaluates the fairness of one of the three original dimensions of evaluation in terms of those who interact with/or consume the final recommendations [23]. In contrast, P-fairness evaluates this from the perspective of those who provide or create the content to be recommended for consumption [23]. In later work, Burke and Abdollahpouri defined two new definitions of multi-stakeholder fairness [22]. They introduced CP-fairness to address the need to evaluate consumer and provider fairness at the same time [22]. Additionally, they presented the definition of

side-stakeholder fairness, which evaluates fairness in terms of stakeholders beyond consumers or providers [2]. It is important to note that evaluating fairness within the recommendation system goes beyond evaluating if a bias exists. Instead, it evaluates if the outcomes of this bias are "fair" or equally distributed across individuals or groups of stakeholders.

Understanding how to define and evaluate fairness and bias of recommendation systems has quickly grown into a seminal area of information retrieval research. Various types of bias related to recommendation systems have been defined and studied in academic and industry settings. Researchers often target studying bias relating to harms of allocation, unequal distribution of exposure, or attention of recommendations within the system [32]. Allocative, or distributional, harms have been studied by evaluating and mitigating biases such as popularity, exposure, ranking (or pairwise), and gender bias [3, 5, 4, 40, 28, 59, 34]. A recent literature review by Ekstrand et al. notes that representational harms can also be studied in recommendation systems but focuses on representation in terms of the provider and how stakeholders view their distribution within the system, not their numerical representation as vector outputs of a recommendation system [32].

There are three overarching methods for mitigating for fairness and bias in RecSys: pre-processing, in-processing, and post-processing [38]. Pre- and post-processing mitigation methods tend to be model agnostic, meaning the technique can be applied in a system regardless of the algorithm (albeit with some restrictions depending on the training and final output data). Pre-processing is implemented on the training and test data leveraged to train the recommendation algorithm [38]. Post-processing techniques mitigate bias after recommendations have been made [38]. Commonly, post-processing is implemented as a re-ranking

technique. In-processing mitigation is often a model intrinsic method developed as a specific algorithm or introduced into the training as a part of the optimization function, most often as a regularization term [38].

In the following passages, we will present various definitions of fairness or bias presented in recommendation system research.

2.2.1 Distributional Harms

Abdollahpouri et al. states that popularity bias is present when "popular items are recommended even more frequently than their popularity would warrant" and demonstrated the propensity for it to occur in recommendation systems in various works [3, 3, 5, 4]. Geyik et al.'s audit of ranking on LinkedIn is an example of evaluating the *rank fairness* in a recommendation system in production. For example, [59] evaluated popularity bias to measure the fairness of the distribution of content binned by original popularity rank.

[60] and [35], evaluated the fairness of distribution of gender in music recommendations. [34] evaluated the distribution of author gender in book recommendations when using different recommendation techniques. [40]'s audit of ranking on LinkedIn is an example of evaluating the *rank fairness* in a recommendation system in production. [25] evaluated how gender affects *rank performance* on job sites, such as Indeed and Monster. [28] conducted a large scale audit on *ranking bias* for Amazon recommendations. More recently, [32] produced an extensive literature review to help frame the current environment for recommendation fairness. [32]'s review provides guidance in using exposure and pairwise fairness within recommendation fairness evaluation frameworks. The review also notes that there is still a great need for research on developing best practices

for when and how to leverage recommendation metrics. More detailed information on multi-stakeholder, pairwise, and exposure fairness can be found in [32]'s literature review.

2.2.2 Set Membership Bias

Set membership bias is primarily concerned with the distribution fairness within the final recommendation set. *Top-k and set-based fairness* focus on the evaluation of bias in a ranked recommendation set membership[40, 91, 94]. Set-based fairness is the quantification of representation of protected groups relative to a comparison distribution[91]. Yang et al. presented three metrics for measuring set-based fairness: normalized discounted difference, normalized discounted KL divergence, and normalized discounted ratio[91]. These measures are calculated by computing, such as proportion difference and KL divergence, at discrete points within the ranked set, then compound the values with a logarithmic discount[91]. The measurements are discounted to capture the importance of rank[91]. Zehr et al. built upon this notion with top-k fairness. This version of fairness in ranking evaluates a set of k candidates in a ranked list for containing a required proportion of group membership[94]. They defined this notion as *ranked group fairness* [94]. Geyik et al. presented two metrics for measuring top-k fairness in an industry setting at LinkedIn[40]. They introduced Skew@k and a cumulative version of normalized discounted KL divergence. Skew@k calculates the logarithmic ratio of the proportion of recommendations having a specific attribute among the top-k ranked results against a corresponding desired proportion of that attribute[40]. Their version of KL divergence reflects the weighted average of Skew@k overall defined attribute values[40].

Another version of set membership bias is *equity of attention*. Individual equity of attention is an individual fairness metric that evaluates bias in set membership[17]. This metric aims to evaluate the diversity of content groups represented in a recommendation set. Unlike top-k or set-based fairness, it does not compare against a predefined group distribution. Mehrotra et al. presented a similar metric in a group evaluation setting which they called group fairness[59]. To avoid confusion with the popular definition of group fairness mentioned above, we refer to Mehrotra et al.'s metric as group equity of attention.

Tsintzou et al. introduced *bias disparity* in recommendation systems to measures the relative change in bias value between the training and recommendation sets for a user[85]. Bias is calculated as the conditional probability of a provider group being recommended or selected given the user group[85]. This probability is calculated with preference ratios, the fraction of chosen or recommended content from the provider group, compared to the probability of choosing the provider group at random[85].

2.2.3 Rank Position Bias

Metrics for measuring bias in rank position leverage the idea of pairwise comparison to compare group rank performance[14, 48]. Beutel et al. leverage pairwise comparisons to create the notion of pairwise fairness[14]. Pairwise fairness consists of three definitions: pairwise fairness, intra-group pairwise fairness, and inter-group pairwise fairness[14]. Each of these definitions is evaluated with a version of pairwise accuracy. This metric calculates "the probability that a clicked item is ranked above another relevant unclicked item"[14]. Pairwise fairness evaluates pairwise accuracy across groups for equality[14]. Intra-group pair-

wise fairness is evaluated with intra-group pairwise accuracy, which is calculated by comparing the clicked item with other relevant unclicked items in the same group[14]. Inter-group pairwise fairness leverages inter-group pairwise accuracy, which compares the clicked item with other items not in its group[14]. Beutel et al. formulated pairwise fairness to enable comparisons of relationships between different groups when ranked in the same set.

Kuhlman et al. leveraged pairwise comparisons to introduce rank parity. They leverage the calculation of concordant and discordant pairs to form the basis of their metrics[48]. The concordant and discordant pairs are determined by comparing predicted ranks to true rankings for the set[48]. They introduce three rank parity metrics: rank equality error, rank calibration error, and rank parity error[48]. Rank equality error was created to capture the notion of equalized odds for rankings by capturing how often items from one group are incorrectly ranked higher than another group[48]. Rank calibration error captures the overall error made for items in a group. It is based on the classic fairness notion of calibration[48]. Rank parity error reflects the notion of statistical parity by measuring how often one group is favored over the other regardless of their positions in the true ranking[48].

2.2.4 Rating Bias

Yao et al. presented the idea of non-parity metrics to calculate fairness in collaborative filtering systems[92]. The goal of non-parity metrics was to provide a way to measure bias without access to ground truth ratings in collaborative filtering systems[92]. Instead of comparing binary accuracy, their set of non-parity metrics calculates the difference in predicted values between groups[92].

Yao et al. introduced four metrics for this type of measurement. Value unfairness is formulated to measure the difference in signed estimation error between user groups[92]. Absolute unfairness is similar to value unfairness but measures the difference in absolute estimation error[92]. Underestimation and overestimation unfairness measures differences in how predictions underestimate or overestimate true ratings for users[92].

2.2.5 Exposure Bias

Singh and Joachims presented the idea of evaluating fairness by comparing exposure resulting from rank predictions between groups[77]. Their method of evaluating exposure fairness requires defining a way to measure the amount of exposure resulting from a specific rank position. They compute exposure from the utility (or relevance) and position of the item in a ranked list[77]. Item exposure is summed over all possible items in a group to determine the predicted exposure of a producer group[77].

2.3 Latent Bias in Natural Language Processing

Our proposed evaluation framework for measuring attribute association bias in latent factor recommendation model embeddings is inspired by natural language processing (NLP) methods that attempt to measure binary gender bias in word embeddings. These methods looked to understand associations between gender-neutral words (like "scientist" or "nurse") and words indicative of a specific gender (like "man" or "woman"). Past work has identified gender biases in pretrained static word embeddings, contextual word embeddings from large language models, and embeddings of larger linguistic units like sentences [20, 24, 12, 83, 98, 16, 57,

52]. Because pretrained word & sentence embeddings are widely used as input for many NLP models, there is potential for biases in embeddings to be propagated or amplified in downstream text classification and generation tasks [98, 68].

Pretrained word embeddings are a fundamental input to many natural language processing (NLP) tasks. As word embeddings are understood to capture robust syntactic and semantic meaning, it follows that undesirable social biases have been found encoded in these linguistic representations [24, 49]. Undesirable stereotypes and biases have been identified in both static and contextual embeddings and at both the word and sentence level of embeddings [20, 12, 98, 83, 52]. This is an important area for research, as Zhao et al. have shown that biases in embeddings can be surfaced or amplified in downstream tasks.

Despite the non-linear models used to train word embeddings, semantic and syntactic information can be understood through linear relations (Mikolov, 2013). The geometric properties of embeddings can be represented through analogies such as queen : king :: man : ?, where simple linear arithmetic can be used to complete the analogy with woman. Early research in embedding bias used these gendered analogies as evidence for undesirable gendered meaning captured in embeddings [24, 98, 20].

Various methods for evaluation and measurement were introduced to better understand this bias for natural language processing. In a seminal work by Bolukbasi et al., the difference vectors of pairs of words that have are definitionally gendered (such as she-he or woman-man) are used to establish a gender subspace [20]. This logic for identifying a gender subspace has since been replicated across a range of NLP embedding contexts and research [36, 16, 82, 56].

Caliskan et al. introduced the Word Embedding Association Test (WEAT) as a method for evaluating bias in text that was inspired by the Implicit Association

TEST (IAT) used by psychologists [24] . In the case of gender bias, WEAT uses cosine similarity to compare a set of target words that are definitionally gender neutral (such as science) with sets of gendered attribute words (such as male, and female). Another measure for bias, defined by Ethayarajh et al., is Relational Inner Product Association (RIPA) [36]. This method uses a relation vector, which is created through definitionally gendered word sets, and then takes the inner product between the relation vector and any word vector. RIPA builds from the work

RIPA and WEAT were designed as methods of quantifying semantic gender in text. Another approach to identifying gendered meaning in word embeddings has focused on grammatical gender in languages such as Spanish and French. Instead of applying PCA to learn a gender direction, linear classifiers such as Linear Discriminant Analysis (LDA) or Support Vector Classifiers (SVC) have been used to classify grammatical gender [101, 67]. In this approach, the model is trained to classify grammatically female and masculine words, such that a decision hyperplane can be used towards signal disentanglement.

This dissertation looks to these seminal works in gender bias evaluation in natural language embeddings to formulate novel approaches for recommendations. The evaluation metrics and methods we introduce for recommendations can be seen as novel due to the inherent differences between recommendation embeddings and language embeddings. One main difference we account for is the fact that definitionally gendered or stereotyped one-to-one relationships between entities do not exist in recommendations settings. Additionally, item entities often do not have defined sensitive attributes, meaning that the practitioner may need to investigate how attribute association bias occurs within their system against the user entities that define the sensitive attribute within the space.

Chapter 3

RecSys Auditing Frameworks in Practice

The methodologies we share in this dissertation are intentionally designed to support implementation in practice, specifically within broader algorithmic auditing frameworks. Broad system-level frameworks provide essential structure to approaching the ambiguous problem space of auditing for harms and bias in industrial systems [70, 13]. [70] introduced the seminal auditing framework of "SMACTR", standing for "Scoping, Mapping, Artifact Collection, Testing, and Reflection." SMACTR is viewed as a framework for a large-scale audit, explicitly accounting for "procedures and documentation, as well as considering system outputs" [11]. Our proposed framework of methodologies fall more in line with the idea of "disaggregated evaluations," which is targeted by the framework, SIIM, introduced in this section. Disaggregated evaluations may also be captured by the "Testing" step of SMACTR, where the focus lies on analyzing AI outputs for harm or bias [70, 13, 11].

Our attribute association evaluation and mitigation methodology framework

provides practitioners with guidance to analyze attribute association bias systematically with the SIIM framework presented and discussed below. We do this by specifically focusing on how to scope the attribute to be measured for bias, identifying and implementing our methodologies to determine the existence of bias, and then finally, flagging for significant results and significant changes in bias after mitigation. The following section and subsections were presented and published by the author at the conference RecSys in 2022.

3.1 Algorithmic Auditing in Practice

In 2019, EU reports called the development of industry standards 2-4 years away [37]. Around the same time, Jobin et al. found more than 80 documents containing ethical principles or guidelines for AI, also pointing to the need for more implementation guidance rather than principles alone [45]. Raji et al. began to address this need by introducing SMACTR to provide a framework for a wider auditing context [71]. Bakalar et al. added to this body of work concerning responsible AI in practice by presenting insight and guidance into implementing fairness for binary decisions in industry systems [9]. However, their best practices may not apply to more complicated workflows such as recommendation systems.

A growing selection of tools has been created to attempt to bridge this gap, such as Fairness360 and Fairlearn [18]. However, such tools are not always applicable to an auditing team. Many of these tools are generally tailored to classification or regression tasks rather than ranking techniques used in standard recommendation settings [74]. This leaves practitioners evaluating recommendations with the overwhelming tasks of translating, implementing, and standardizing research for evaluating fairness in their systems.

Measuring fairness comes with challenges, such as interpretation issues for non-experts [74], the volume of conflicting fairness metrics [62], and a steep learning curve for the average practitioner [51]. In addition to these challenges, as Jacobs and Wallach point out, fairness is essentially a contested construct [44]. Different definitions of fairness should be seen not as only different measurement operationalizations, but rather as different conceptualizations or perspectives on values. Determining which algorithmic harms are to be evaluated in a measurement exercise depends on such choices [10]. Helpful end-to-end frameworks around algorithmic auditing can provide high-level organizational anchoring (e.g. [71]); however, these frameworks provide little guidance on detailed challenges that occur when starting to audit a specific product area. Organizations may have to retrofit a consistent approach across hundreds of existing systems, while accounting for system-specific nuances, making one-off studies unscaleable. These challenges lead to a frustrating situation where teams enthusiastically begin auditing systems, but the lack of industry standard tooling or guidance halts their progress.

3.2 SIIM Framework for Auditing in Practice

The framework we introduce, SIIM, consists of four steps: scope, identify, implement, and monitor and flag. The first step, scope, addresses the problem of determining "what" to analyze. For example, what sensitive attribute should be analyzed in our LFR model vector output? The second step, identify, focuses on determining the best-suited methodologies for said analysis based on the scope and outputs of the system. The third step, implement, represents the time dedicated to conducting the analysis and determining how to manipulate the data to

leverage identified methods. Finally, the fourth step, monitor and flag, answers the vital question of if significant levels of bias exist within the system.

This framework was designed at Spotify to address challenges encountered during a central algorithmic auditing effort, and after specific products have committed to auditing their products. SIIM is designed to help practitioners develop organized and strategic audits for evaluating bias and harms in their recommendation systems. Below, we will introduce the steps of the SIIM framework as well as highlight key challenges practitioners may face when trying to implement the step in practice. By sharing these challenges, we hope to provide inspiration for future research directions to lower the barrier to reducing harms in production-level recommendation systems.

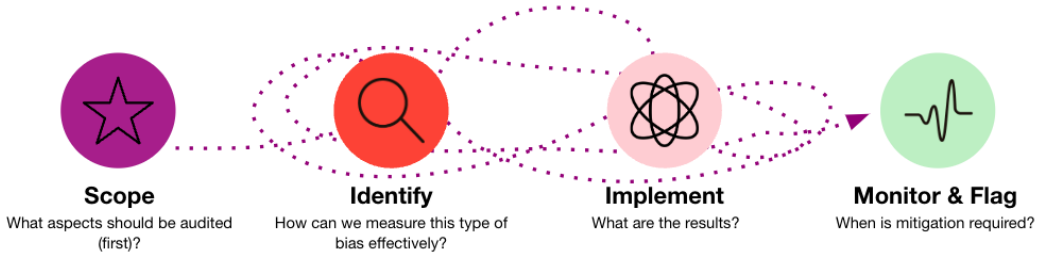


Figure 3.1: The SIIM framework breaks down the auditing process into four key steps that seek to answer essential questions encountered while auditing a recommendation system in practice.

3.3 Steps in quantitative evaluations

The SIIM framework can be organized into four distinct steps: **s**cope, **i**dentify, **i**mplement, and **m**onitor and flag. We organized our steps into these four groups based on four targeted questions we needed to answer to conduct our audit:

- **S**cope: What aspects should be audited (first)?

- **I**dentify: How can we measure this type of bias effectively?
- **I**mplement: What are the results?
- **M**onitor and flag: When is mitigation required?

Figure X illustrates the intended steps in order: scope, identify, implement, and monitor & flag. However, in practice, the journey is non-linear with teams possibly revisiting steps multiple times through out the process of auditing their systems. For example, after implementing a specific evaluation or mitigation method, the practitioner may found that it did not exactly satisfy the scope of their audit, thus requiring them to revisit the identify step in order to find a more well-suited method for their specific scenario. In figure Y, we provide a highlevel overview of possible sub-steps within SIIM. Scope tends to be the most qualitative step, focused on scoping the audit and defining requirements for implementing fairness or evaluating bias within their system. The next step, Identify, is where the research begins with teams identifying possible methods for satisfying the scope of their audit. Once those methods have been identified for testing, practitioners can Implement the audit. Implementation may include engineering or sampling the dataset, technically implementing identified methodologies from research, and testing the methods. This step also includes the first iteration of the audit. If all goes well and implementation has satisfied the original scope of the audit, the practitioner can move to the Monitor and flag step, which focuses on identifying if mitigation is needed, testing and implementing mitigation stratigies, and finally setting thresholds for future flagging for intervention.

In the sections below, we provide a high-level overview of challenges we encountered with each step. (Our presentation will include specific examples of these challenges that were not included in this proposal due to length constraints.)

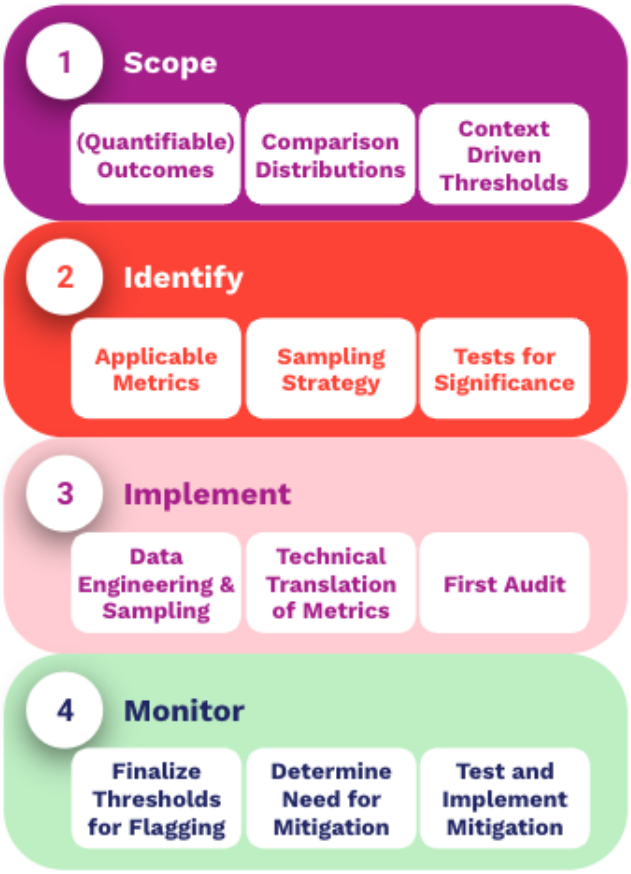


Figure 3.2: Each step of the SIIM framework consists of sub-steps to address when auditing one's recommendation system.

3.3.1 Scope

The first step to evaluating bias is to scope what exactly needs to be measured. Larger organizations may have provided policies and frameworks of topics of concerns (e.g., legal requirements around data restrictions, societal concerns around popularity bias, gender and racial equity, etc.), but that does not immediately translate into exact playbooks that guide where to start and how to scope an evaluation; this still requires judgement and agreement on who should be involved in this judgement. The challenge for practitioners is determining when their scope is *good enough* to conduct an initial evaluation, not necessarily if it is optimal.

Determining and prioritizing focus is especially difficult in a multi-stakeholder recommendation system; evaluating consumer versus provider fairness are two separate objectives, which may require different and (or) contrasting approaches to evaluation. Differing internal and external pressures, as well as practicalities such as data availability and quality, need planning and prioritization. In addition to determining focus, practitioners need to understand where in the system, or rather system of systems, to measure. Industry recommendation systems can be complex, consisting of building blocks across many teams. Even when a practitioner may know what type of bias needs to be evaluated, it may be difficult to translate the evaluation techniques company-wide across systems, due to different algorithms, system design, and available data — at the same time this scoping can have serious consequences.

3.3.2 Identify

Once the evaluation has been scoped, the practitioner needs to determine how to measure the targeted bias. Finding the correct measurement technique not only includes sifting through research, but also testing if the metrics are compatible with the system(s) tested. Even after testing, we ended up with an extensive list of possible metrics to evaluate consumer and producer fairness. Filtering this list was difficult akin to the discussion in [44]. This turned a seemingly simple task of choosing a metric into a complex debate between alternate constructs of recommendation fairness [44], while product leaders need concrete and concise guidance. Choosing metrics wrongly can have grave consequences, but this gap gives practitioners the sizeable task of narrowing down potential research metrics. This means that larger organizations with resources to do this investigation have

an advantage. In addition to finding the correct metrics, the practitioner is also often tasked with explaining them to stakeholders that need to be involved for feedback. Depending on the stakeholder, choosing a complicated fairness metric could introduce more confusion, reducing motivation in teams who want to show results. We found that many of the metrics were not intuitive when presented to an audience with little to no experience with standard recommendation or ranking fairness metrics. In addition, even for expert audiences, a communication gap exists. Impactful wins from an organizational algorithmic responsibility perspective, such as being able to monitor a platform across systems, may not be top of mind to individual teams motivated to show results for their own system.

3.3.3 Implement

At this step, practitioners are essentially "ready" to begin measuring the targeted bias. Unfortunately, due to the lack of tooling available for recommendations and ranking, completing this step requires technical expertise to translate chosen research methods into code for future evaluation. Implementing a technical translation of the research can be challenging for practitioner teams if they do not have adequate access to engineering support. Even if they do, calculated recommendation fairness metrics at scale can be complex because they may require calculation for each user recommendation set.

Technical implementation also involves making decisions that can appear to be trivial, but in reality have potential to heavily impact results. In our experience, choosing comparison distribution and sampling methods can introduce unforeseen difficulties. For example, it is possible to evaluate different definitions of fairness depending on how the comparison distribution is defined [40]. Choosing the

wrong comparison distribution can introduce more bias in itself [88]. Guidance is necessary to help practitioners design their comparison distributions effectively without possibly reinforcing the harmful effects of their recommendation system. Scoping a new sampling strategy incurs similar challenges. The list of alternate ways to the above approaches makes it easy to get bogged down in analysis of all possible options, but a team may feel they need to "start somewhere" for initial analysis to learn and move towards an impactful audit. However, it can be difficult for practitioners to definitively know that they are headed in the right direction, and get the space to pivot and return to this step when they are not. This means that feedback and evaluation moments have to be built in, requiring buy-in from both central algorithmic responsibility researchers and involved teams.

3.3.4 Monitor and flag

Developing a method and a metric to monitor is not enough to audit a system for algorithmic harm. Decisions need to be made of when to act, whom to alert, and what follow up actions are required across different teams. Determining thresholds to flag bias for mitigation is another challenging task in our evaluation. There is a difference between "practical" and "statistical" significance when it comes to assessing meaningful harm [65]. Statistical significance is influenced by many decisions (e.g., alpha, one- vs. two-sided test) that are generally designed to assess whether results occurred due to chance, not whether the results were meaningful. "Significant" doesn't mean practical, and "marginally better than what was before" might not be very impactful.

Determining "practical" and context-driven cut-offs beyond traditional "sta-

tistical" thresholds requires making difficult decisions about what the company deems "fair". Creating these concrete goals and cut-offs, and deciding when to act, is a substantial undertaking. Given the lack of standards, practitioners may have to develop their own unique guidelines to develop context-driven thresholds. This process can be taxing because of the potential to prioritize sub-optimally, to introduce or ignore bias, or not really "making a difference" (see also [35] for related examples on creator gender representation in recommendations). Practitioners making these difficult decisions could have potentially large downstream effects on creators and users, without tools to further investigate their impact. Assessing the "practical" significance often requires subject matter expertise that even a trained fairness practitioner often does not have. Without concrete guidance and inter-team agreements and protocols, it is hard for teams to know when they are required to take action even if their systems are monitored.

Part II

Auditing Attribute Association Bias

Chapter 4

Scope: Research Setting

This dissertation will explore AAB in four different datasets. Three of these datasets are publicly available with the fourth reflecting proprietary Spotify podcast data. The Spotify dataset was available for research through my employment and individual industry PhD agreement with Spotify to produce and publish research pertaining to this specific model and its input and output datasets. For all scenarios, we evaluate AAB in terms of the candidate pool generating component of a hybrid recommendation system. This is due to our focus on latent factor recommendation algorithms primarily being used to produce candidate pools, not final ranked results in a recommendation system.

We chose to explore various types of biases in unrelated datasets to determine how well our proposed methodologies for evaluating and mitigating AAB function under differing circumstances. Additionally, these experiments enable us to observe how different biases may be stronger or weaker under certain conditions, such as the types of algorithms leveraged or if feedback loops were simulated. This dissertation examines gender, size, and age bias. Each of these biases have the ability to create stereotypical experiences and have been examined in both

academic and industry settings. Details behind our choice in bias will be discussed within the scenario specific sections below.

4.1 Spotify Podcast Recommendation

We evaluate user gender AAB regarding podcast genre by implementing our framework on an industry hybrid recommendation system. We decided to evaluate the first component of the system as the earliest point in which this type of bias could be introduced into final recommendations. This component is an industry production-level candidate generation model for podcast recommendation; it uses an LFR algorithm to create pools of podcast vectors by user to be ranked for final recommendation lists. More specifically, candidates are generated via a deep neural network (DNN) recommendation model, a setup commonly used in industry systems [63, 26]. It mimics matrix factorization collaborative filtering via a DNN and is trained with candidate sampling and importance weighting to account for potential popularity bias. Model inputs include user features, podcast ids, and binary labels representing positive or negative implicit feedback. Final user and podcast (or item) representation embedding vectors are collected as outputs from the model.

We trained the model with and without using user gender as a feature to understand the counterfactual effects and potential for implicit bias when trained without explicit use of the sensitive feature. This also enabled us to explore use of our framework for creating baselines for mitigation methods, such as removing sensitive attributes from a model. All analysis and training were conducted offline due to the sensitivity of mitigating user gender bias in an online industry system.

4.1.1 Gender Bias in Podcast Preferences

Our case study provides a quantitative deep dive into biased listening behavior previously observed in qualitative academic studies. These studies provided guidance for grouping our entity vectors to implement our analysis methods; we decided that exploring user gender AAB would provide a good case study for experimenting with our framework and contributing novel insight into this area of research. For example, [19] found listeners of true crime podcasts were predominately female and showed three specific motivations. [27] found that motivations for podcast use in young adults did not significantly change across gender but across genres, signaling a potential change in gender and genre combined. [80] presented results showing that various demographic parameters, including gender, drove podcast interests in Latina/o/x young people. However, these stereotypes have yet to be researched quantitatively via bias evaluation in the context of recommendation systems. In this case study, we continue investigating the relationship between gender and genre by analyzing and quantifying potential user gender bias captured in the trained latent space representation from a recommendation model.

Defining Entity Sets

After deciding to target user gender AAB, we needed to determine how to group our entities for analysis. We can group male and female users into our attribute-defining entity sets to target user gender. Nevertheless, we must also determine how to frame our entity sets to test for user gender bias given the variety of possible stereotypes associated with user gender. For example, one could frame the analysis to understand if there is user gender AAB regarding creator gender.

An Edison report focusing on podcast listening behavior of women found that women would listen to more podcasts if there were more female hosts within the podcast space [50]. Another industry study by AT&T found that men were likelier to listen to podcasts hosted by men [1].

Instead of framing our analysis that way, we chose to target how user gender may become associated with specific *genres* of podcasts. We looked to past research on podcast genre listening behaviors by gender to determine which genres we should define as test entity groups. In particular, [19] noted that true crime podcast listeners are more likely to be female than male, and true crime is one of the most popular genres in female listening [50]. In contrast, sports podcasts have been found to have a primarily male listenership [42]. When observing proprietary data concerning gender share in listenership, we confirmed that these two genres were significantly skewed towards women for true crime and towards men for sports. Given these findings, we explored gender AAB for podcasts labeled as true crime or sports.

Our podcast vectors were labeled by predetermined podcast genres. These genre labels were defined via self-selection from podcast hosts and behind-the-scenes cataloging of podcasts. Because a podcast can be classified under multiple genres, we required podcasts labeled as true crime not to be labeled as sports and vice versa.

Due to past research on binary stereotypes in gender in information retrieval literature, we chose to approach stereotypes in a binary sense between feminine and masculine genres [34, 69] In respect to genre categories, we leverage predetermined genre specifications for podcasts. These genres are attributed to podcasts via self-selection from podcast hosts as well as behind the scenes cataloguing of podcasts. In all, there are 21 genres available for analysis and podcasts can be

classified under multiple genres. Given past research, we focus on differences in bias between true crime and other genres. Additionally, we create groups of genres based on historical gendered listening: stereotypically male, female, or neutral. A genre is assigned to a specific gender group if more than 65% of historical listeners are of that gender. At the time of this research, the female genre group consisted of Health & Fitness, True Crime, and Kids & Family while the male genre group consisted of Sports, Technology, and Leisure.

4.1.2 Production-level Candidate Pool Generation Evaluation

We evaluate user gender AAB regarding podcast genre by implementing our framework on an industry hybrid recommendation system. We decided to evaluate the first component of the system as the earliest point in which this type of bias could be introduced into final recommendations. This component is an industry production-level candidate generation model for podcast recommendation; it uses an LFR algorithm to create pools of podcast vectors by user to be ranked for final recommendation lists. More specifically, candidates are generated via a deep neural network (DNN) recommendation model, a setup commonly used in industry systems [63, 26]. It mimics matrix factorization collaborative filtering via a DNN and is trained with candidate sampling and importance weighting to account for potential popularity bias. Model inputs include user features, podcast ids, and binary labels representing positive or negative implicit feedback. Final user and podcast (or item) representation embedding vectors are collected as outputs from the model.

We trained the model with and without using user gender as a feature to

understand the counterfactual effects and potential for implicit bias when trained without explicit use of the sensitive feature. This also enabled us to explore use of our framework for creating baselines for mitigation methods, such as removing sensitive attributes from a model. All analysis and training were conducted offline due to the sensitivity of mitigating user gender bias in an online industry system.

4.1.3 Experimentation Settings

We created our evaluation data set by randomly sampling 9,500 female and male users to create a final set of 19,000 users. Our podcast vector data set comprised 31,181 English podcasts from the DNN recommendation model. We restricted our analysis of recommendations to users registered in the United States and podcasts created by English speakers. We chose this subset of data to minimize the possibility of location and language confounds, which could potentially affect gender bias measurements due to differences in cultural norms. In the future, it would be interesting to research how gender stereotypes are found as algorithmic bias differently in recommendations concerning the location and language of the users and served content.

Podcast listening has also been shown to demonstrate the potential for gender bias. For example, [19] found listeners of true crime podcasts were predominately female and showed three specific motivations. [27] found that motivations did not significantly change across gender but did change across genres, signaling a potential change in gender and genre combined. [80] presented results showing that podcast interests in Latina/o/xs young people were driven by various demographic parameters, including gender. In this paper, we continue investigating the relationship between user gender and genre by analyzing how user gender in-

fluences podcast recommendations through the entanglement of user gender and trained item vector representations. We evaluate user gender bias in terms of female and male users.

4.2 MovieLens Movie Recommendation

Our setting for experimentation differs from that presented for the Spotify case study due to there being less constraints on experimentation with the public datasets. Given this, we investigated our proposed definition of bias with a popular public dataset to explore if this phenomenon can be present beyond our original scenario. We leverage the MovieLens dataset, a commonly used public dataset for recommendation research, as one of our case studies. We do this by setting non-data specific experimentation scenarios to analyze how AAB may differ according to the recommendation algorithm leveraged and the recommendation scenario. In the following subsections, we provide an overview for these general experimentation settings as well as define dataset specific details.

4.2.1 Gender Bias in Movie Preferences

Similar to podcast listening, gender bias has also been studied regarding movie preferences. This bias has been studied quantitatively by evaluating gender biased language in screenplays [72]. Ramakrishna et al. found that romantic and comedy movies consisted of more feminine language than non-romantic or non-comedy movies[72]. Additionally, they found that the opposite held true for Action and Crime movies[72].

Stereotypes define romance, comedy and "melodramatic" movies as movies for women. But for men, action and horror are supposed to reign supreme. These

stereotypes have been confirmed in qualitative studies observing gender preference for genres [66]. More recently, less well-known gender stereotypes were confirmed qualitatively in a study observing the accuracy of gender stereotypes on actual movie-genre preferences [89]. Wuhr et al. found that gender stereotypes do exist, however, they over estimated the actual size in difference between gender preferences [89]. This study looked at 17 different genres and found that drama and romance were more preferred by women[89]. Action, adventure, erotic, fantasy, horror, mystery, science fiction, war, and Western were found to be more preferred by men[89]. Only six genres were considered equally popular amongst men and women: comedy, animation, crime, heimat, history and thriller[89].

4.2.2 Age Bias in Movie Preferences

Movie genres are often associated with particular age groups. For instance, action and adventure films are commonly perceived as targeting a younger, predominantly male audience, while romantic comedies may be seen as more suitable for older viewers or women. Age was found to be a helpful predictor and machine learning feature for movie genre preference [86]. In a study of Italian movie watchers between 18-65 found that younger participants showed more preference, via regression analysis, for romance, drama, horror, thriller, action and sci-fi [43]. However, this quantitative finding was different from that found by where horror was preferred more by younger individuals and drama was more preferred by an older audience [47]. Similar to exploring user gender bias, we will look at user age bias and how it can be quantified within the latent space.

4.2.3 Experimentation Settings

We use two samples of data, first, MovieLens-100k which consists of 943 users, 1682 movies, and 100,000 interactions. The other sample is for 1M interactions with x users, x movies, and 10000029 interactions. The dataset includes gender, age, occupation and location as user features, as well as, genres for the movies. For our evaluation, we implement our framework for observing user gender and user age bias on the movie genres.

When reviewing the results for the public dataset, MovieLens, we will focus first on observing how AAB manifests itself when leveraging a deep neural network. This allows us to remain slightly more in line with our experimentation settings defined for the Spotify case study. With the MovieLens dataset, we provide insight into bias directions and bias metrics based on the algorithms used, and where appropriate how data sampling can significantly change results. Additionally, we will provide a review of how AAB differed across the three modeling scenarios targeting the evaluation of gender bias in the latent recommendation space.

We leverage the python package RecBole in order to use pre-built and evaluated recommenders for collaborative filtering scenarios. As mentioned before, we leverage DMF as our model for initial evaluation of AAB across our three cases studies. We train the data off of the interaction matrix, not the rating matrix.

We leverage the following type of latent factor recommendation models available in said package: Bayesian Personalized Ranking (BPR), Deep Matrix Factorization (DMF), and Neural Collaborative Filtering (NCF). We chose these models based on their frequent use in research to examine recommendations for both explicit and implicit feedback. Additionally, these three models represent

three different ways of approaching recommendation algorithms, thus allowing us to explore when our proposed methods can or cannot be used. BPR and DMF both inherently model relationships between the user and item entities by leveraging a final layer calculating the cosine similarity between the user and item. BPR differs from DMF by implementing methods to model ranking, while DMF does not take ranking into account. NCF is fundamentally different from both of these methods because it does not explicitly model relationships via cosine similarity or the dot product. Additionally, the entity embeddings are defined earlier in the algorithmic architecture meaning it is farther removed from the final recommendation prediction. By observing NCF in addition to BPR and DMF, we are able to evaluate how this algorithmic difference affects the final trained latent space. We explore the ramifications on this primarily in section X, finding that our methods are not appropriate for NCF.

For our experiments, we leverage the network and parameters as proposed and implemented in RecBole via their pre-defined set-ups for each model. These parameters are set to train up to 300 epochs with potential to stop training after 10 steps. Weights are not decayed after each epoch. Recbole leverages ADAM as the optimizer with a learning rate of 0.001. The training batch size is set to 2048 records with an evaluation batch size of 4096. The embedding size for all entities across all possible datasets is 64. The data is split between training, testing, and validation at 80, 10, and 10 percent.

Chapter 5

Identify: Methodology Framework

This chapter focuses on presenting the methodologies we leverage for addressing AAB in practice. When conducting an audit of AAB, we look to understand three main attributes of AAB before mitigation. First, we need to understand if AAB *exists* within the trained latent space. Second, we must determine the level of *significance* of the bias for the given scenario. Finally, we look to examine the potential for *amplification* of the bias. These three attributes form the basis of our AAB evaluation framework, ESA: existence, significance, and amplification.

Evaluating each of these AAB attributes are specific to the implementation step of the SIIM framework. Our first three sections address implementation instructions for the ESA framework. Our final section, focuses on implementation steps required for mitigating AAB. The mitigation section captures the final step of the SIIM framework, Monitor & flag. This is due to the fact that we need to *implement* an evaluation and then *flag* a model for mitigation.

The first three sections introduce and provide implementation instructions

for exploring AAB with our ESA framework. In addition to introducing these methods, we provide guidance for identifying which methods to use during the evaluation to help practitioners find the correct methods for their specific use case. The methods we introduce are designed to guide the practitioner from initial exploration of the AAB to the final determination of which groups should be targeted for mitigation.

First, we evaluate the existence of AAB by designing and calculating bias directions. Second, we determine the significance of the bias for the scoped scenario by identifying significant entity groups and calculating bias evaluation metrics for further analysis. Third, we analyze potential for bias amplification by providing guidance for evaluating feedback loops and potential reinforcing bias. We introduce these methods in order of the type of analysis the practitioner wishes to conduct, from initial exploration to targeted measurement for determining mitigation needs. Thus, we provide support for evaluating bias across different phases of analysis while addressing bias and harm within one's recommendation system.

When evaluating bias amplification, we evaluate feedback loops, where bias is reinforced over time, with simulations of recommendations. These simulations mimic a situation where AAB is left unchecked and recommendation models are retrained on historical data from the potentially biased predictions. We also look at reinforcing bias via the ability for entity embeddings to relay sensitive information and bias as features in downstream models. We conduct this study by leveraging classification models trained on the entity embeddings.

Finally, we present methods for mitigation of AAB in a recommendation setting. We provide an overview of possible approaches and present our final choices for mitigation. We look at all three possible mitigation steps: pre-processing, post-processing, and intrinsic.

5.1 Introducing the ESA Framework

As previously mentioned, we look to understand the *existence*, *significance*, and *amplification* of the AAB. When first exploring the *existence* of AAB, we suggest exploring the significance of scoped attribute bias directions. Finding significant bias directions in the latent recommendation space may alert practitioners to the existence of significant levels of AAB. Bias directions provide quantitative means for determining if a significant relationship exists between the entities and the attribute. One can leverage these methods to answer the question: "does AAB exist?". In addition to bias directions, we showcase how one can visualize the attribute space. We do not recommend this method as a necessary evaluation step due to it's unreliability for useful output across differing data domains.

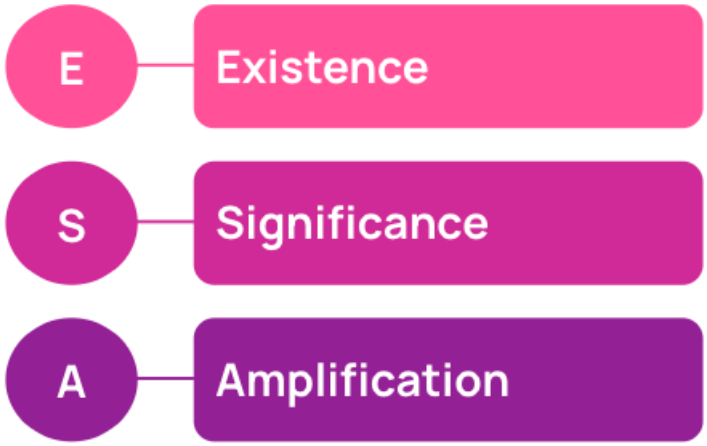


Figure 5.1: The ESA framework consists of three steps to help practitioners understand attribute association bias in their systems.

If a clear attribute relationship has been found to exist within the latent space, one may investigate the problem further by measuring the level of *significance* of the bias present. In many cases, the practitioner may not know which group in

a multi-group setting is experiencing more or less significant AAB. We provide a series of steps to help remedy this issue. Our next step focuses on identifying significant groups for further analysis when there are multiple possible groupings of entities. This step is helpful when distinct binary relationships between groups have not been defined by the practitioner.

One will also want to conduct level setting for future mitigation and direct measurement by implementing AAB evaluation metrics to test for significant levels of AAB and create statistical baselines for evaluating if mitigations are successful. Practitioners can implement these methods to address the question: "how strong is the level of AAB in my system?". Between identifying groups for analysis and calculating bias metrics, we can determine the *significance* of the bias within the system.

Finally, we address the *amplification* of the bias by evaluating how the AAB may change over time if left unchecked via recommendation simulation. Additionally, we explore how the latent space can capture systematic bias by leveraging classification scenarios to evaluate the levels of stereotyping bias captured by the embeddings. These classification scenarios can be leveraged to understand amplification of information if these embeddings are used in downstream models or recommendation components.

When describing implementation details, we refer to the attribute-defining entity sets as A and B. Each entity, $a \in A$ and $b \in B$, is assigned a binary label representing the attribute, with the label of set A being one and that of B being zero, or vice versa. These two entity sets can be considered opposing if their labeled attribute is mutually exclusive. Entity sets used to test for AAB will be referred to as E and P. It is assumed that one entity set is hypothesized to show heavier stereotyping towards one of the opposing attribute entity sets.

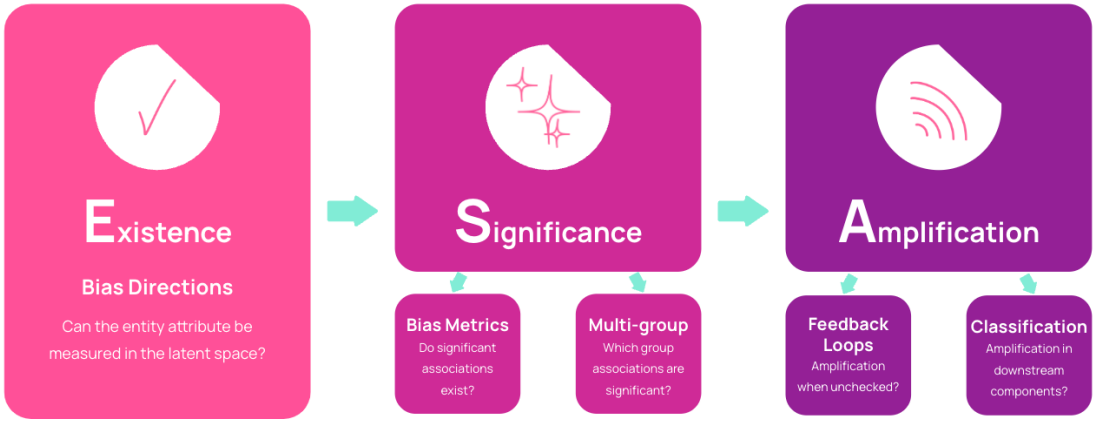


Figure 5.2: The ESA framework consists of specific methodologies to evaluate attribute association bias to address each of the three framework steps.

5.1.1 Existence: Bias Directions

Calculating Binary Attribute Bias Directions

Calculating attribute bias directions can serve as another method for exploring the existence of bias in one's system. These *AAB direction vectors* represent how the attribute is distinguished as a vector direction between A and B within the trained latent space. These vectors can be used for: (a) exploring individual entities by identifying users or items whose embeddings have high similarity with a particular AAB vector for further examination; (b) comparing recommendation systems by using the bias vectors to calculate association bias metrics for each system; and (c) exploring classification scenarios.

We present three methods for computing AAB direction vectors: centroid difference, SVC vector direction, and PCA vector direction. Unlike related work in NLP association bias research (e.g., [20]), the centroid difference and SVC vector direction calculations do not require practitioners to have distinct representation embedding pairings between entities in A and B, making them suitable for recommendation systems and data.

Centroid Difference The simplest method for computing an attribute's association bias vector is to take the difference between the centroid of A and the centroid of B (also referred to as attribute vector mapping [54]). This method is best used for capturing differences in *average* attribute behavior. The centroid method is the most readily interpretable of the three due to its simple calculation and thus serves as a good starting point for exploring the attribute space. However, it is essential to note that this method tends to be more conservative in estimating bias due to variance being averaged out in the process. It may not adequately capture significant nuances in behavior within the space, and other direction techniques may be required to reflect more complex attribute bias behavior.

SVC Vector Direction Our second approach computes the association bias vector using parameters from a linear support classification (SVC) model trained to predict the attribute. We draw inspiration for this technique from past NLP research that trained SVC models to predict grammatical gender in word embeddings [67, 101]. The entity vectors and labels in sets A and B are used as training data for the SVC model. The attribute bias direction is created from the final attribute layer of the model to capture the subspace representing significant attribute meaning. The selection and assignment of entities to A and B can substantially impact the computed bias direction; in our case study, we compare bias directions computed on random samples of users versus most stereotypically-gendered ones. This direction methodology is best used to capture more distinct nuances of attribute bias that may be lost when entity vectors are averaged in the centroid difference method.

PCA Vector Direction The final method calculates the bias direction by using the general methodology introduced by [20], which is based on conducting principal component analysis (PCA) on parallel attribute pair vectors. The final attribute bias direction is the first eigenvector of these vectors, capturing the majority of the variance found describing the group of vector pairs. Similar to the methods above, two groups of opposing vectors must be defined to create the final pairing of vectors. However, unlike the two methods above, implementing PCA for vector direction creation requires distinct attribute pair vectors. This better enables visualization for transparency, but should only be used if the practitioner is confident in their entity pairings for defining the attribute. Therefore, this may not be a good starting point for bias exploration. We show this caveat in our case study by presenting the downfalls of randomly selecting attribute entity pairs for creating a PCA vector direction.

Testing for Significance

Unlike testing significance for metrics, testing the significance of a latent direction requires validating that the direction captures attribute behavior. We aim to determine that the direction is not capturing a random relationship between entity vectors but a distinct attribute-related relationship. We suggest the following comparisons for significance testing:

- *Cosine similarities between the bias direction and entities in opposing entity sets A and B*: This test determines if the two sets have significantly different relationships with the bias direction.
- *Cosine similarity between the bias direction and entities in A and B versus the entities' cosine similarity with a randomly-sampled bias direction*

vector: This test examines if the entity sets have a statistically significant relationship with the calculated bias direction versus a random direction.

- *Cosine similarity of entity vectors from A (or B) with the bias direction and that of random vectors with the bias direction*: This test specifies that the entity has a significant relationship with the bias direction in comparison to random entities and the direction, further validating that the relationship between the entities and computed bias direction is not random.

All three tests must show statistical significance for one to determine that the bias direction captures a non-random relationship between the two attribute-defining entity sets and the calculated bias directions. Given the number of statistical tests conducted, one may wish to leverage a p-value with the Bonferroni correction or other techniques to account for multiple tests for significance [76].

Calculating Multi-categorical Bias Directions

Exploring latent AAB for a multi-categorical attribute group raises more challenges during the bias evaluation than that for simple binary attribute scenarios. From our experimentation, we found that it is best practice to understand the pairwise binary relationships between groups of one's attribute to help guide the analysis. However, if one is concerned with the multi-categorical attribute bias present in the latent space, we provide a work around framework for exploring a multi-categorical attribute situation from scratch. For this dissertation, we propose methods of finding multi-categorical defining directions without needing a non-linear model or deep learning solution. Future iterations of this work could explore more complicated implementations which have been showcased to have success in discovering complicated latent directions via adversarial networks

for GAN models (Discovering Interpretable Latent Space Directions of GANs Beyond Binary Attributes). It is important to note that these linear based direction creation methods will not be able to capture the same level of nuance as a more complicated direction discovery implementation. These methods may serve well for flagging some existence of AAB, but future analysis and mitigation may require more nuanced and complicated implementations to insure that the correct level of bias is reflected by the surfaced attribute direction.

We present the following two definitions for calculating multi-categorical bias directions:

- Holistic Attribute Bias
- Singular Group Attribute Bias

Singular Group Attribute bias reflects the same type of bias directions described previously, however, these directions have to be made for each group or in a pairwise fashion. There are two ways one can implement this, one versus many or one versus one, when creating the direction. In a one versus one scenario, there would be a direction for each group pairing in a pairwise fashion. This would also be accompanied by pairwise analysis. For one versus many pairings, the bias directions would be created by setting one entity group to one category and the other entity group to reflect all other entities.

Holistic Attribute bias investigates the existence of an attribute direction that captures behavior of all the possible groups within the latent space. One can capture holistic attribute bias by measuring the variance between the singular group attribute bias directions (either as one-to-one or one-to-many). The variance vector can be calculated by taking the first principal component of the singular

group attribute bias (SGAB) directions. These SGAB directions can be calculated leveraging any of the bias direction methods mentioned above, but the practitioner may wish to leverage SVC bias directions due to their greater ability to capture group nuances during training. Additionally, SVC models can be trained for multi-categorical classification, allowing us to use the train and test accuracy as another way to evaluate the significance of the directions. If the SVC model is unable to perform adequately, one can hypothesize that holistic attribute bias directions may not be the best path forward for AAB evaluation.

In the case of leveraging SVC bias directions, one can train a one-to-one or one-to-many linear support vector machine model. Next, the first principal component can be calculated on the linear coefficient vectors from the SVC model. This final principal component vector represents the most variance between the predictive hyperplane of the various groups. When comparing entities with this final holistic attribute vector, we can assume that entities with high absolute values of cosine similarity with the vector are more affected by the variance between these groups. Additionally, groups with opposite levels of cosine similarities can be classified as groups with highly different behaviors with respect to the holistic attribute. In the results section, we explore holistic attribute bias directions with the MovieLens data for user age AAB. We find that holistic bias can be extremely difficult to capture in this simplistic implementation, particularly for independent sub-groups. Holistic attribute bias directions may be best used when the groups reflect scales of one category, such as age or temperature. When confronted with independent multi-categorical groups, it may be best to calculate the singular attribute bias directions in a pairwise fashion to understand which groups show more or less group-specific behaviors within their latent space.

Testing for significance

Similar to the binary attribute direction, holistic attribute directions require testing for significance as well. However, these directions require extra testing to ensure that the attribute direction is unique to the targeted attribute, this requires significance testing against other attribute independent defining groups. We propose replacing the randomly generated binary vector tests with another series of tests for evaluating the significance of the direction.

It is important to note that not every group will result in significant behaviors according to the holistic attribute direction. This means that those groups are not highly associated with the variance defining the holistic attribute. In that situation, one may wish to reiterate on the direction leveraging only the entities that are significantly associated with the holistic attribute direction to produce a more accurate direction for capturing the targeted attribute.

Significance testing for multi-categorical holistic AAB directions should include:

- *Entity cosine similarities between the multi-categorical bias direction versus an independent attribute bias direction*: This test determines the bias direction is significantly different from another attribute defining bias direction. An example of this would be comparing the cosine similarities of the entity vectors with an age bias direction and gender bias direction. Comparing directions that may be interdependent will result in less accurate results. In the case that all attributes are heavily correlated, one should test for significance only leveraging a randomly genderated direction (as described below).

- *Entity cosine similarities between the multi-categorical bias direction versus*

a randomly generated direction: This test examines if the multi-categorical bias direction captures significant and non-random behavior within the latent space.

- *Cosine similarity of group entity vectors with the bias direction and that of an unrelated attribute bias direction*: This test specifies that the entity has a significant relationship with the bias direction in comparison to it's relationship with an unrelated attribute bias direction. This test helps highlight which groups are showing more or less significant relationships with the resulting multi-categorical bias direction.

- *Cosine similarity of group entity vectors with the bias direction and that of a randomly generated direction*: This test investigates similar significance behavior as the previous test but targets understanding if specific groups have non-random relationships with the multi-categorical bias direction.

5.1.2 Significance: Bias Metrics & Multi-Group Evaluation

We propose two metric methods for capturing the significance of AAB culminating from LFR algorithms. Two NLP techniques inspire these evaluation methods for evaluating latent gender bias in word embeddings: Word Embedding Association Test (WEAT [24]) and Relational Inner Product Association (RIPA [36]). We chose these methods based on their acceptance within the NLP community as reliable metrics for measuring bias in vector representations of words [31, 95].

We also address a current problematic limitation of recommendation evaluation presented by academic literature: the lack of guidance surrounding explo-

ration of multivariate scenarios. In general, algorithmic bias research is done on binary groups because it is inherently easier to measure biased relationships between two groups rather than multiple groups. In many cases, the sensitive attribute is not binary, thus making the methods above not possible for analyzing the bias. This section will also detail possible methodologies for analyzing bias in a multi-group setting according to the test entity groups. Multi-group or multi-categorical techniques for the attribute defining group (such as male and female users for defining gender) are detailed in the bias directions and metrics sections. For identifying significant test entity groups and relationships, we detail a framework for surfacing specific groups of the test entities for further bias analysis leveraging bias metrics or classification scenarios.

Entity Binary Attribute Association Metrics & Test

This set of metrics, inspired by WEAT [24], can be used to understand how AAB manifests in user-user and user-item comparisons by computing vector similarity between entities of interest and members of the two attribute groups defined previously (A and B). These metrics require two sets of users or items to evaluate the AAB (E and P), where one entity set is hypothesized to show heavier stereotyping than the other. There are three interrelated metrics: entity attribute association (EAA), group entity attribute association (GEAA), and differential entity attribute association (DEAA).

EAA measures the AAB for a single entity $\varepsilon \in \{E \cup P\}$, calculated as the difference in mean cosine similarity of ε to attribute entities in A and B. Positive EAA scores represent a higher association with attribute A while negative scores

signal higher association with attribute B:

$$EAA(\varepsilon, A, B) = \frac{\sum_{a \in A} cos(\varepsilon, a)}{|A|} - \frac{\sum_{b \in B} cos(\varepsilon, b)}{|B|} \tag{5.1}$$

GEAA is the sum of all entity attribute association scores for a set of entities (E or P):

$$GEAA(E, A, B) = \sum_{\varepsilon \in E} EAA(\varepsilon, A, B) \tag{5.2}$$

Finally, DEAA acts as the test statistic for permutation testing by measuring the scale in the difference between the GEAA of E and P. Positive DEAA scores signal that entities in E show more association with attribute A than entities in P and vice versa for negative scores:

$$DEAA(E, P, A, B) = GEAA(E, A, B) - GEAA(P, A, B) \tag{5.3}$$

We use permutation testing to evaluate if there is a significant difference in how the test entity sets relate to the attribute entity sets. Additionally, we adopt the calculation for effect size presented by [24] to evaluate the normalized separation between EAA distributions of the test entity sets:

$$\frac{\frac{GEAA(E,A,B))}{|E|} - \frac{GEAA(P,A,B)}{|P|}}{stddev(EAA(E \bigcup P, A, B))} \tag{5.4}$$

Recommendation Relational Inner Product Association (R-RIPA)

We also provide a metric, R-RIPA, that is similar to the prior metrics, but parameterized by a user-defined attribute bias direction. This provides more flexibility for the practitioner to use computed AAB vectors based on SVC and PCA, or

other user-defined attribute directions in general. Additionally, this metric may be more robust to fluctuations or outliers that can affect metrics heavily reliant on group averages over entities. Unlike EAA metrics, R-RIPA can be used in a non-binary fashion by calculating R-RIPA against a multicategorical bias direction.

We base this metric on RIPA [36], which is calculated with a relation vector representing the first principal component of the difference between word pairings in an attribute-defining set. We modify RIPA to require a relation vector ψ that represents a user-defined attribute bias direction between A and B. R-RIPA for an entity set E is computed as:

$$R\text{-}RIPA(E, \psi) = \frac{\sum_{\varepsilon \in E} cos(\varepsilon, \psi)}{|E|} \tag{5.5}$$

The effect size for R-RIPA between entity sets E and P is then:

$$\frac{R\text{-}RIPA(E, \psi) - R\text{-}RIPA(P, \psi)}{stddev\{cos(\varepsilon, \psi) \mid E \bigcup P\}} \tag{5.6}$$

Bias Metrics: Testing for Significance Permutation tests can be used to determine the significance of the AAB evaluation metrics [76, 24]. When evaluating EAA metrics, one can test for the significance of the entity-specific metric, GEAA, and the entity-difference metric, DEAA. A significant GEAA means a biased relationship exists between the entity group and the defined attribute. A significant DEAA represents a significant difference in the level of AAB between the two entity test sets. For R-RIPA, permutation tests can be used to determine if the difference between the entity set association attribute bias is significant. One can test for a significant difference in AAB between entity test

sets by comparing the two populations' cosine similarity scores with the bias direction leveraged when calculating R-RIPA. An alternative method is to calculate R-RIPA for smaller samples of the test entity sets and apply the Wilcoxon rank-sum test or similar [76].

Surfacing Significant Test Entity Groups from Multiple Groups

Capturing multiple groups accurately via one cumulative metric is inherently difficult thus the need for group to group analysis to identify significant relationships. Knowing this, a current work around is implementing pairwise metrics, or divergence metrics, to identify the existence of skew. Unfortunately, the computation of pairwise metrics for a large number of groups quickly becomes an expensive and slow process. Problems arise for divergence metrics as well due to the inability to quickly pinpoint which group is causing the skew without binary comparisons. As a result, guidance for multi-group scenarios still requires binary group analysis.

We propose a two-step framework for exploring and identifying possible group attributes to help practitioners quickly target which groups should be further explored for AAB. Our framework focuses on identifying singular groups showcasing signs of higher levels of AAB. The first step focuses on preparing the data for evaluation, where we frame our problem by defining the groups as "features" which may potentially cause an effect on a bias direction based metric. The second step looks to identify "feature" importance when predicting the bias direction based metric.

Preparing the Data As detailed in the bias metric section, one of the most important ways to measure AAB is through calculating the cosine similarity

between the attribute entities or attribute bias direction and the test entities. The base for the group test statistics are single entity calculations of association bias. We can leverage these single entity calculations to determine which group attributes should be used for future binary comparisons when calculating the DEAA and R-RIPA. These entity calculations will function as our target when evaluating feature importance and causality.

Exploring Feature Importance We can find the feature importance of test entity categories by modeling the bias direction based metric against the categorical features. After training this model, we can look to the categorical feature importance to guide our evaluation, allowing us to pinpoint which test entity groups should be audited for AAB. There are a variety of modeling and feature importance strategies, but in this dissertation, we leverage multiple linear regression and decision trees to explore multi-categorical feature importance on AAB metrics. These two models provide insight into the feature importance as well as the relationship with the bias attribute.

5.1.3 Amplification: Feedback Loops & Classification

The ability for recommendation models to capture and exploit sensitive attributes is a well known phenomenon. This has lead to the popularity of disentanglement mitigation and adversarial learning for mitigating bias in the latent space. Even though these research areas are growing in popularity, many of the papers lack content concerning the evaluation of the association bias captured in the space. These papers tend to focus on performance metrics or fairness metrics, but do not evaluate how the latent space has changed pre- and post- mitigation. With these steps of evaluation, we look to explore how AAB may change over time if left

unchecked. Additionally, we explore how the latent space can capture systematic bias by leveraging classification scenarios to evaluate the levels of stereotyping bias captured by the embeddings.

Simulating Feedback Loops

Feedback loops and bias have been proven to be influenced by popularity bias Commonly, the propensity for feedback loops has been observed based on simulation experiments. These types of tests are performed offline when online testing is unavailable, which is often the case in academic settings, additionally testing for compounding harm in an online setting may not be responsible when offline simulation is available. The method we choose for simulations is inspired by Mansoury et al. where they simulate recommendations by "iteratively generating recommendation lists to the users and updating their profile by adding the selected items from those recommendation lists based on an acceptance probability. In addition to modeling the acceptance probability, we also chose randomly from the final top-k groups of content.

We explore how AAB can change over time if left unchecked or unmitigated. Our goal is to pinpoint how feedback loops and AAB are interrelated. We do this by testing if there are significant increases in the bias metrics after a series of simulations. If there is a significant increase, for our tested scenario, we can confirm that over time feedback loops are created through reinforcement of stereotypes entanglement. Each simulation consists of a training iteration of the recommendation model.

It is important to note that each iteration created a new model, and there was no warm start training between iterations. We simulated recommendations for five iterations. We created the new set of interactions by calculating the top

10 potential movie recommendations and then narrowing the pool to five movies with an acceptance probability based on the rank of the prediction. We modeled the acceptance probability based on the equation from Abdollapouri et al. Each subsequent retraining was for 50 epochs. We chose to reduce the number of epochs based on the first model having the best epoch at 35. In addition to analyzing differences in the bias directions, we analyzed changes in bias metrics for our highlighted stereotypical groupings of genres. Differences were analyzed between the original entity vectors and the final simulation (round five) entity vectors. Our goal was to determine if, left unchecked, user gender AAB would increase significantly over time. The following sections will reflect the results for one of the three recommendation algorithms. Finally, we will compare results to determine if specific algorithms pose more risk to developing potentially harmful feedback loops over time.

Our work differs from previous feedback loop simulations focusing on popularity bias because we wish to observe how sensitive semantic attributes strengthen, weaken, or remain static in the latent space over time. Showcasing that AAB can compound if left unchecked supports the hypothesis that recommendations are capable of deeply influencing our everyday interactions with content. Previous work has showed that diversity decreases with feedback loops, but we wish to observe how stereotypical experiences may increase for a user, potentially creating harmful situations such as radicalization and rabbit holing.

We implement this type of evaluation only for the MovieLens dataset since this work was out of scope for the proprietary podcast audit.

Testing for Significance When testing changes in a feedback loop, we found that it's important to not only test changes in the group statistics, but the indi-

vidual statistics as well. In some cases, nuanced changes in the individual entity vectors may become lost in the aggregate metric calculations, resulting in looking over key insights in how the latent space changes over time. In order to account for this, we not only test changes in the aggregate bias metrics, we also test for differences in the individual bias metrics as well. As noted in section, the aggregate bias metrics reflect individual calculations, EAA for DEAA, and the cosine similarity between the entity and bias vector for R-RIPA. We test changes between the iterations by calculating the Kolmogorov-Smirnov test. We test for a null hypothesis specific to the "sign" of the attribute direction. For example, if "female" entities are reflected with a negative direction than we test for if the simulated entity individual metrics are less than the original entity individual metrics.

Exploring Reinforcing Bias with Classification Models

Past NLP research has used classification models to show how heavily word embeddings capture bias and demonstrate possible downstream effects, particularly along the lines of binary gender [41, 12]. This implementation builds upon that research by evaluating entity embeddings for systematic bias and their risk for introducing bias in downstream models. We conduct this by leveraging the entity embeddings for various classification scenarios where entity embeddings are leveraged as the training features. We propose training a classifier on user or item embeddings and their target attribute, meaning the model is trained on entity sets A and B and their associated attribute labels. This classifier can then be leveraged to explore how attribute bias and stereotypes are captured within the trained latent space, e.g., by comparing predictions of new entities not in A and B. This method is especially advantageous when assessing the potential for

amplifying representation harm when using item or user embeddings in models downstream from the original recommendation system.

We leverage three classification scenarios when evaluating potential for reinforcing bias between the datasets. These scenarios were framed to understand the ability for item entity vectors to convey sensitive attributes as features in downstream models. First, we look at attribute bias by item group. This looks to understand if attribute bias changes in relation to the attribute percentage historically interacting with a specific item group. Next, we model the user's attribute from their item interaction history. This scenario explores the ability for item history to convey a sensitive attribute, which could create privacy concerns depending on the attribute leveraged. Finally, we predict the attribute of the items and observe if there are certain genres that are classified more often as one attribute group versus the other.

It is important to note that we test how results change across simulations for only the MovieLens dataset since temporal evaluation was out of scope for the proprietary podcast audit.

Testing for Significance Analyzing the results of classification scenarios should account for how the scenario was scoped and the goal of the bias evaluation. For example, suppose a practitioner is analyzing the potential for AAB in a downstream model that uses learned item embeddings from a recommendation system. In that case, they should first measure overall accuracy to determine if the embeddings relay the attribute correctly. If the practitioner is also concerned with unfair levels of AAB across items in specific stereotype groups, one could leverage classification fairness metrics to compare performance across specified groups, such as demographic parity or equalized odds [58]. We refer to [58] for

an overview of classification fairness or bias metrics in such scenarios. Differences in metrics should be tested with widely accepted statistical testing methods for classification models, such as a t-test [30].

5.2 Mitigating AAB

Given AAB is being defined and presented by this dissertation, there is a lack of research directly targeting the mitigation of this specific type of bias. Even though this research may not exist, we can look to NLP research on association bias and recommendation research concerning disentangling bias from the latent space. In this chapter, we explore possible mitigation methods and choose methods for future implementation. We provide distinct reasons for these choices to showcase how one can distinguish between methods in practice. It is important to note that mitigation was not formally conducted for the Spotify Podcast Recommendation case study due to the specific goals for the proprietary project. However, our choice to remove user gender from training could be seen as a pre-processing mitigation technique. Results of this pre-processing mitigation are captured along with overall metric results in chapter six since the decision to remove gender was made primarily to understand the effect of gender on AAB in the recommendation model. Mitigation methods presented in the following sections are framed according to their implementation on the MovieLens dataset.

5.2.1 Pre-Processing Mitigation

We experiment with pre-processing in two separate ways for our use cases. First, for the podcast case study, we remove gender as a user feature during the training of the model. We chose to implement this simple method since it was implemented

in an industry setting. Within this setting, we were testing the effects of implementing a new company policy requiring the removal of sensitive features during the training of our models. The case study we evaluate presents the implications of following that policy.

For the MovieLens case study, we chose to resample the data given the heavy skew towards male users in the original training data. In these cases of attribute skew, it is common to resample data to achieve a more balanced training distribution (CITE). There are a variety of ways one can sample data for training, we chose to focus on the more simplistic techniques of over or under sampling the data for model training. We conducted both over and under sampling to account for the potential of losing recommendation quality when not leveraging the entirity of the dataset. When oversampling, we randomly sampled interactions from female users and added these sampled interactions to the training dataset to create a balanced number of interactions across male and female users. The undersampled dataset was created by randomly removing male user interactions from the training dataset to achieve balance. The oversampled dataset consisted of 1507538 interactions and the undersampled dataset contained 492880 interactions. These datasets were used to train a BPR recommendation model with the same parameters as the original recommendation model.

5.2.2 Post-Processing Mitigation

Hard debiasing (Olukbasi) requires a way to identify "neutral" entities and entity pairings. This requirement is difficult to achieve in a RecSys environment due to the inherent "neutrality" of item entities and subjective nature of defining entity pairings to capture a group attribute (unless counterfactual training is available

within the system). Additionally, the entities leveraged to define the attribute may not have a true "neutral" grouping. For example, what is considered a "neutral" age for a user? Practically, there may not be a quantifiable answer to that question. In order to answer that question, one would have to define what "neutral" means in accordance to the entities and attribute they are evaluating within their system. Creating this subjective definition poses a risk of introducing more bias in the evaluation and mitigation. In order to avoid falling into the trap of quantifying a subjective definition, we look to for inspiration in mitigation methods that do not require these types of distinct pairings or neutral definitions, or at least do not rely heavily on them to complete the mitigation.

In particular, we look to work presented by Ravfogel and Dev which look to remove or disentangle trained neural representations of potentially harmful concepts and attribute associations. In comparison to other works, their proposed methods do not heavily rely on word pairings and are more easily adapted to other domains beyond NLP and word embeddings. In this dissertation, we explore leveraging these methods for reducing binary AAB in recommendation entity embeddings. We found that the two most popular use cases for this in NLP research was iterative nullspace projection and orthogonal subspace correction and rectification (OSCaR).

OSCaR requires two entity based directions to unbias the latent space vectors. In the case for recommendation systems, this would translate to needing a specific user and item bias direction relating to the sensitive attribute, such as male versus female and true crime versus sports. By requiring this second direction, it places another binary direction requirement on the mitigation which would make it more computationally expensive when one of the entity groupings is not binary (such as genre) or if the comparison entity groups are not known. For example,

with our movie recommendation case study, this type of mitigation would need to be implemented multiple times to account for the various relationships such as the genre directions romance to action or children's to action. We chose to explore iterative nullspace projection instead of pursuing OSCaR based methods for recommendation bias mitigation. We came to this decision since that method focuses on removing the stereotype based on the defining entity (such as user to gender or user to age), not the specific relationship between two entity directions. This specification makes the method more flexible to recommendation mitigation goals that may transcend specific user to item defined stereotypes. In the following subsection, we will describe iterative null space projection and how we use this mitigation method for the MovieLens case study.

Iterative null space projection mitigation consists of repeated training of linear classifiers to predict the sensitive attribute for futural removal from the latent space. This method achieves this by projecting the entity vectors onto the final nullspace of these iteratively trained linear classifiers. By leveraging SVC models (as used for finding our bias directions), we can leverage this method to remove the linear separation and dependence between the AAB direction and the entity vectors. We approach this in two ways, first by removing only the originally created AAB direction and second by implementing this iteratively to account for the ability for future classifiers being capable of capturing linear separation for our binary sensitive attribute. We assess the success of these methods based on two goals of removing AAB while minimally affecting the original end task of producing accurate recommendations. We test this by comparing original attribute association and accuracy metric results with the post-mitigation results for statistically significant differences.

5.2.3 Intrinsic Mitigation

We explore the use of adversarial machine learning as a potential approach in mitigating AAB in latent factor recommendation algorithms. Adversarial recommendation has primarily been leveraged to create robust recommendation systems by increasing the security to algorithmic attacks and, more recently, generative recommendations. Another less explored application has been leveraging adversarial components to increase the privacy of stakeholders within the recommendation system. As we show in section Y, one can leverage recommendation lists to predict sensitive features of consumers within the recommendation system. This ability creates privacy risks for owner's of the recommendation systems which could potentially lead to legal ramifications (SOURCE ALGO BOOK). Beigi et al introduced the idea of Recommendation with Attribute Protection, which "simultaneously recommends relevant items and counters private-attribute inference attacks". They alter the BPR algorithm to include an adversarial component to decrease the ability for an attacker to learn sensitive information from a consumer's recommendation lists. This adversarial component is fed a concatenated vector of the user and item entity vectors to predict the user's sensitive attribute information. It is trained by maximizing the loss of this component, resulting in obscuring sensitive attribute information from the latent entity vectors.

Given this training goal, adversarial recommendation for obscuring private attributes would be a good candidate for mitigating AAB since it should theoretically remove stereotyped relationships within the trained latent space. We test it's ability to mitigate AAB by adding an adversarial component to our original BPR algorithm which looks to predict the gender of the user from the concatenated user and item entity vectors. We test the success of the method similar

to the post-processing mitigation by testing bias and performance metrics for significant difference.

Chapter 6

Implement: Existence

This chapter marks the implementation of the first step of the ESA framework, which looks to understand the existence of AAB within the trained latent recommendation space. We leverage the bias direction methods previously presented to explore AAB in our two case studies. For the Spotify podcast recommendation case study, we determine the significance of the user gender bias direction calculated from our sample of male and female podcast listeners. We observe two different possible bias directions for the MovieLens movie recommendation case study. The first bias direction targets the significance of binary user gender. Our second implementation explores our ability to leverage presented methods for identifying significant multi-categorical bias directions. We explore this by creating a bias direction for user age.

6.1 Spotify Podcast Recommendation

This section presents results from implementing our framework and flagging the existence or change in bias when removing user gender as a feature. The goal of

this evaluation was to test whether our methodologies were successful in capturing and understanding user gender AAB from our case study's model. We do not explore our methods for identifying significant groups in a multi-group scenario due to our distinct choice to target the relationship between true crime and sports podcasts in relation to user gender.

Before calculating bias directions and metrics, we visualized our embeddings. Figures 6.1 & 6.2 show these latent space visualizations of user and podcast embeddings with and without user gender as a model feature respectively. In both cases, our first principal component had a high cosine similarity to the centroid difference (0.89 and 0.91, with gender and without gender as a feature, respectively). We observed both user clusters and podcast clusters emerge along the first principal component for the embeddings trained with and without user gender as a feature. When we projected the entity embeddings trained without gender, we observed a similar pattern of clustering as in the case of podcasts trained with gender.

Both the user and podcast embeddings trained without gender have flipped directionality for the gender direction due to being trained separately from the with-gender embeddings. However, we observe the same relationship of gendered clustering for users along the first principal component in both contexts. Our projections demonstrate that user and podcast embeddings trained without a gender vector still have latent gendered meanings encoded. In the case of podcasts, we observe a weaker separation along this axis, although it is still possible that feature clusters could be derived.

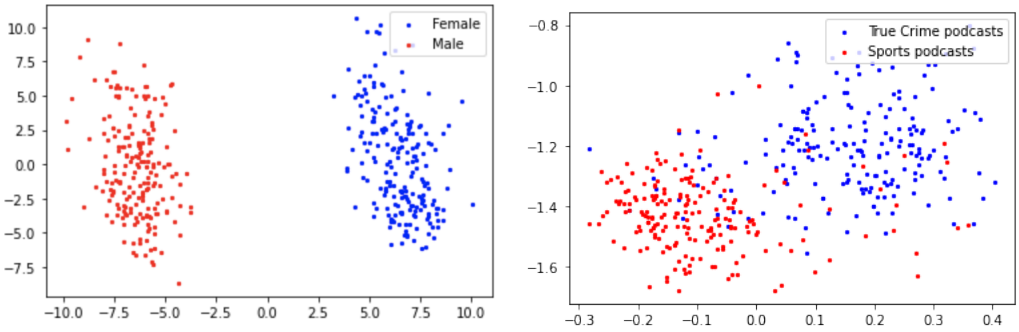


Figure 6.1: Projection of user embeddings along the first and second PCA components of the 400 most biased users trained *with gender* (left). Podcasts trained in the same embedding space also show clusters along the same principal components (right).

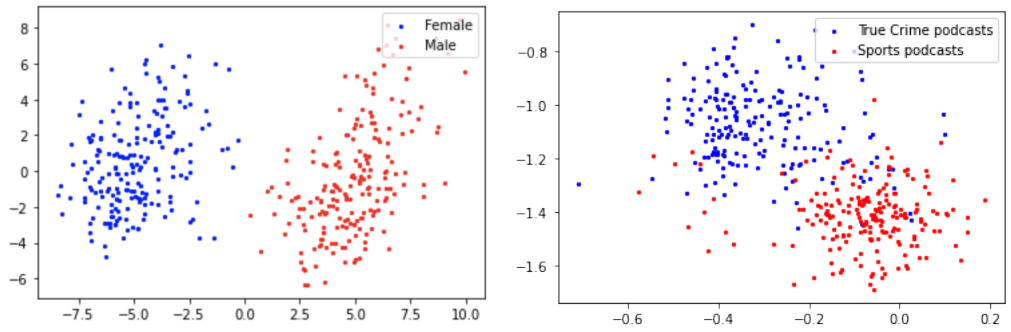


Figure 6.2: Projection of user embeddings along the first and second PCA components of the 400 most biased users trained *without gender* (left). Podcasts trained in the same embedding space also show clusters along the same principal components (right).

6.1.1 Bias Directions

When testing the resulting bias directions for significance, we found that all possible direction methods, except for the PCA bias direction, resulted in significant results for our three recommended statistical tests. With the Bonferroni correction, statistical tests were considered significant if p ¡ 0.0033. Beyond the PCA direction, bias directions were significant regardless of if user gender was or was not an explicit attribute used during model training. However, numeric test statistics were greater for our statistical tests when user gender was present

during training.

We found that the PCA direction created from random pairings between male and female user vectors failed the test comparing the cosine similarities between the entity and bias direction versus that of random vectors and the bias direction. Since the other two tests were significant, this pointed to the bias direction capturing a significant difference between the attribute entity groups and a significant direction in the space but not a significant relationship between the entity attribute and the bias direction. This result means that the bias direction should not be used for further analysis of gender AAB in the space, since it could easily capture other attribute behavior within the latent space and thus result in inaccurate observations of AAB. It is essential to statistically test one's bias directions since acting on inaccurate results could inadvertently introduce more harm instead of lowering harm in subsequent mitigation.

In addition to our proposed flagging methodology, we evaluated the SVC direction based on its model's test accuracy. We split our data into training and test data on an 80-20 split; the results are in Table 6.1. Our first iteration SVC model trained on a random subset of users achieved 99.3% test accuracy in our with-gender user vectors and 82.2% test accuracy in our non-gender user vectors. Our mixed-method Centroid-SVC direction (CSVC-1) trained on the 200 "most biased" users achieved 96.0% with-gender test accuracy and 73.6% non-gender test accuracy. The other mixed-method CSVC direction (CSVC-2) trained on the 2500 "most biased" users according to the centroid difference gender direction (CD) achieved 96.4% with-gender test accuracy and 81.5% non-gender test accuracy. In order to maintain consistency between test accuracies, we evaluated using the same test set across all trained SVC models.

The significant decrease in test accuracy for CSVC-2 could be attributed to

SVC Training Data		SVC Accuracy	
		Train	Test
SVC	WG	0.992	0.993
	NG	0.829	0.829
CSVC-1	WG	1.0	0.960
	NG	0.959	0.736
CSVC-2	WG	1.0	0.964
	NG	0.943	0.815

Table 6.1: Training and test accuracy of three SVC models used to create bias directions: trained on all users (SVC), the 400 most biased users (CSVC-1), and the 5000 most biased users (CSVC-2). User "bias" was calculated from the cosine similarity with the centroid direction. Accuracy is shown for when user gender was leveraged during training (WG) and then removed as a simple mitigation technique (NG).

multiple factors, such as a reduction in training data, overfitting to training data, or a reduction in gender explainability for all users via the CD. The difference between with-gender and non-gender test accuracy for CSVC-1 is particularly interesting since we intentionally trained the with-gender and non-gender models on the same users. The reduced ability of these 200 users to accurately predict gender when embeddings are not explicitly trained with user gender as a feature signals that the explicit use of this feature does strengthen the significance of gender in user embeddings within the trained latent space. The perfect training accuracy achieved by training our SVC models on the "most" biased users, as found by calculating the cosine similarity with the centroid vector, one can see that the gender centroids do accurately capture gender associations. Additionally, the difference in training and test accuracy for these two models signals that the more stereotypically "gender biased" the user is, the more easily they can be linearly separated by an SVC model. The decrease in training and test accuracy when removing gender as a model attribute shows a reduction, but not complete erasure, in gender attribute association in the latent space.

Finally, the reduced accuracy of the CSVC-1 model with no gender may signal that the resulting bias direction may not be best for calculating bias metrics or analyzing AAB. It would be more prudent to leverage bias directions with high levels of significance via testing and high levels of accuracy (when using SVC to create the bias direction).

We calculated the cosine similarity between each possible direction vector to compare the resulting bias directions. The cosine similarities between the different gender directions reflected in with-gender and non-gender embeddings can be found in Table 6.2. Our comparisons signaled high levels of similarity between all calculated gender directions except for the PCA bias direction. As previously noted, this direction was created by randomly pairing female and male users, finding the difference in their vector directions, and then finding the first eigenvector of their vector differences. Given that this direction is the only one with a low level of cosine similarity, carefully choosing pairs when leveraging this method is essential. We found that randomly pairing users to create a difference vector based on their attribute resulted in our bias direction not accurately capturing potential AAB. The PCA bias direction method should solely be used if the practitioner is confident in their entity pairing methodology to reflect the targeted attribute.

Table 6.2 also shows that the cosine similarity between bias directions from the different methods varied significantly. Even though each direction was significant, this difference demonstrates that each relationship captured is slightly different according to the method used. Additionally, we noticed that these fluctuations decreased when user gender was removed as a model feature. This decrease was expected since user gender was no longer used as a model feature. Additionally, it showcased that the bias directions were capable of relaying implicit, or potentially

Bias Direction		CSVC-1		CSVC-2		CD		PCA	
		WG	NG	WG	NG	WG	NG	WG	NG
SVC	WG	0.71	-	0.75	-	0.79	-	0.03	-
	NG		0.40	-	0.88	-	0.89	-	0.03
CSVC-1	WG			0.99	-	0.98	-	0.08	-
	NG				0.53	-	0.38	-	0.11
CSVC-2	WG					0.99	-	0.06	-
	NG						0.89	-	0.01
CD	WG							0.03	-
	NG								0.06

Table 6.2: Cosine similarity for gender direction vectors created through the following methods: SVC model trained on random sample of 12,000 users (SVC), SVC model trained on 400 most "biased" users (CSVC-1), SVC model trained on 5000 most "biased" users (CSVC-2), centroid difference (CD), and PCA first eigenvector of the difference between 1000 randomly generated male-female vector pairs (PCA). Similarities were calculated for gender direction vectors created from embeddings trained with gender (WG) and without gender (NG).

systematic, bias in the latent space. Given the fluctuations found, we believe it would be responsible for practitioners to explore and test multiple bias directions during analysis to enable more nuanced viewpoints of AAB.

6.2 MovieLens Movie Recommendation

In this section, we will present results from our evaluation of the existence of AAB in the MovieLens 100k and 1M datasets. First, we calculate and evaluate bias directions for six modeling scenarios, BPR, DMF, and NCF trained on the 100k and 1M MovieLens sample. We conduct bias direction evaluation for both binary user gender and multi-categorical user age. It is important to note that we only continue the audit with binary user gender given our primary focus on binary attribute association bias.

	MovieLens100k				
	Recall	MRR	NDCG	Hit	Precision
BPR	0.2085	0.3893	0.2302	0.7402	0.1583
DMF	0.2435	0.4488	0.2210	0.7466	0.1947
NCF	0.2176	0.3787	0.2245	0.7508	0.1532
	MovieLens1M				
BPR	0.1464	0.377	0.2052	0.7267	0.1608
DMF	0.1399	0.3610	0.1943	0.7075	0.1516
NCF	0.1300	0.3439	0.183	0.6919	0.1473

Table 6.3: Here you will find the performance metrics for the three algorithms trained on the MovieLens 100K and 1M sample datasets. All metrics reflect performance for the first 10 recommendations per user.

6.2.1 Bias Directions

In this section, we will review results when calculating and evaluating bias directions created for the MovieLens dataset. We focus on the calculation of bias directions for binary user gender and multi-categorical user age.

Binary User Gender

When evaluating the bias directions for male versus female users in the MovieLens dataset, we found that sample size and algorithm did significantly change bias directions being calculated. BPR and DMF both result in significant gender SVC bias directions. BPR, DMF, and NCF MF embeddings result in significant centroid bias directions. NCF MLP embeddings are the only group not resulting in a significant bias direction when trained on the 1M sample.

We found NCF results to be different from our other two algorithms, which can be attributed to the architecture of the algorithm. Unlike BPR and DMF, there is no interaction between users and items coded into the algorithm. BPR and DMF model user-to-item similarity in the algorithm by including a final co-

sine similarity (or dot product) layer. NCF models interactions via concatenation and element-wise multiplication across embeddings. MLP embeddings contribute to the final prediction via concatenation instead of a dot product or similarity function, as used in BPR, DMF, and the MF layers. Leveraging concatenation would lower the tie between the user and item entities due to optimizing embeddings for the sum of the entities instead of the product of the entities. MF embeddings can capture more of a bias direction compared to MLP embeddings due to the number of abstraction layers between the vectors and the final prediction. This is shown by MLP embeddings from the NCF model being the least accurately separated by its calculated user gender bias direction.

Additionally, the strength of these bias directions increases as the models are trained on more data, which could be due to the imbalance in the datasets between male and female users. When reviewing the bias directions, we find significant directions based on user gender when trained on the 1M dataset for the algorithms BPR, DMF, and the MF component of NCF. Additionally, BPR results in significant directions for the 100k and 1M datasets, allowing us to compare how bias and affected groups differ based on the sample used during training. We can observe that BPR and DMF experience significant bias directions for both the centroid and SVC direction for the 1M dataset but not the 100k dataset. We focus on the 1M dataset given this difference for our future analysis sections. However, in the spirit of experimentation, we explore all significant scenarios to showcase how analysis may differ across datasets and algorithms. Additionally, we continue to explore NCF for the 1M dataset to showcase how AAB may function when it is not significant. This decision is purely to provide more insight into the bias behavior, but for practice, we recommend not proceeding with analysis if the bias directions are not significant.

We can leverage the test statistics to understand better how the male and female user vectors relate to one another and their subsequent bias direction. For example, the DMF embeddings show that female users show more significance when testing against a random direction and random vectors against a bias direction. This signals that female users are more strongly related to the calculated bias direction and each other within the latent space. In the case where gender is more significant in one scenario but not the other, such as that for BPR male users from the 100k set for the centroid direction, it signals that the bias direction strongly captures their behavior. However, the male vector behavior may not be as strongly related. We found that this behavior was accompanied by male users showing lower cosine similarity with the bias direction and the sample distribution being smaller in standard deviation. This behavior could mean that the bias direction does not strongly capture female behavior but that female users show unique behavior concerning the entire population of user entities.

One can also observe how well the bias direction functions by exploring how it separates users. We analyzed the proportion by gender of the 100 most biased users in both directions for our modeling scenarios. If the bias direction functioned well, it should accurately classify more biased users by gender. As shown in Table 6.4 and 6.5, we find that our results mimic our findings when testing for significance. For example, the Centroid direction for the 100k sample resulted in the worst classification results of female users, thus not accurately capturing gender in the latent space.

Interestingly, separation for BPR and DMF were the most similar for the 1M sample, signaling that these models, which inherently model the similarity between users and items, may capture similar levels of bias within their respective trained latent spaces. This extra step in analysis helps confirm that our proposed

significance tests are sufficient in finding significant AAB vectors.

			BPR		DMF		NCF			
							MF		MLP	
			100K	1M	100K	1M	100K	1M	100K	1M
SVC Accuracy	Train		0.7811	0.8007	0.7546	0.7713	0.7904	0.7994	0.7401	0.7208
	Test		0.7195	0.8129	0.7301	0.7466	0.7407	0.7847	0.7037	0.7011
SVC T-Test ABS Value Statistics	M versus F SVC Direction Cosine Similarity		15.88	51.24	14.28	41.32	16.33	48.36	11.28	12.90
	Versus Random Direction	M	9.98	26.30	2.42*	6.88	7.44	5.14	18.44	17.82
	Cosine Similarity	F	6.95	12.45	0.30*	14.91	3.30*	3.62*	3.30*	4.16
	Versus Random Vectors	M	4.49	15.74	5.17	13.72	4.79	14.52	4.25	3.69*
	SVC Direction Cosine Similarity	F	10.19	30.80	8.37	24.63	10.46	29.73	6.84	9.34
Centroid T-Test ABS Value Statistics	M versus F SVC Direction Cosine Similarity		8.83	44.29	9.88	36.01	10.65	40.78	5.93	9.87
	Versus Random Direction	M	10.98	32.55	5.97	8.51	2.49*	6.97	17.15	5.12
	Cosine Similarity	F	4.30	15.20	3.04*	44.74	4.69	13.82	2.01*	1.33*
	Versus Random Vectors	M	2.02*	13.97	2.98*	12.49	3.01	12.11	2.35*	2.34*
	SVC Direction Cosine Similarity	F	6.93	27.33	6.86	22.60	7.62	26.33	3.63*	7.50

Table 6.4: This table reflects the significance of bias directions across our scenarios found during evaluation. First, we look at the train and test accuracy of the SVC bias direction. Second, we showcase the T-test absolute values for our three testing scenarios for the SVC bias direction. Finally, we reflect the same for the Centroid bias direction. We reflect unsignificant results with an asterisk. We can see that only BPR trained on 1M, DMF trained on 1M, and DMF trained on 100k result in significant SVC and Centroid bias directions.

			Accuracy for Top 100 Most Gender Biased Users							
			BPR		DMF		NCF			
							MF		MLP	
	User Gender	Prediction Gender	100k	1m	100k	1m	100k	1m	100k	1m
Centroid	M	M	0.94	0.97	0.91	0.95	0.63	0.95	0.85	0.86
		F	0.06	0.03	0.09	0.05	0.37	0.05	0.15	0.14
	F	M	0.47	0.11	0.47	0.12	0.70	0.11	0.52	0.62
		F	0.53	0.89	0.53	0.88	0.30	0.89	0.48	0.38
SVC	M	M	0.94	0.95	0.98	0.95	0.61	0.94	0.91	0.85
		F	0.06	0.05	0.02	0.05	0.39	0.06	0.09	0.15
	F	M	0.24	0.07	0.32	0.09	0.83	0.09	0.40	0.66
		F	0.76	0.93	0.68	0.91	0.17	0.91	0.60	0.34

Table 6.5: Classification accuracy of the most biased users based on the bias direction. Users are chosen based on their cosine similarity with the bias direction. For example, the 100 most biased female users are the 100 female users with the highest cosine similarity in the female direction. We leverage our SVC classifier to then compare the predictions against the actual gender of the user.

Multi-Categorical User Age

We explore multi-categorical attribute association directions for user age. When calculating age, we experimented with two grouping strategies: three modified age groups and original age categories. Original age categories consisted of seven age groups: under 18, 18-24, 25-34, 35-44, 45-49, 50-55, and over 56, thus resulting in seven categorical hyperplanes for calculating the first principle component. The second strategy grouped users into under 18, 18-49, and over 50. Given the results on binary attribute association directions, we only run this analysis on the MovieLens 1M dataset.

We found a noticeable trend when calculating the cosine similarity with the final age bias direction for the second strategy, with younger groups being more positively associated with the bias direction and a more negative cosine similarity for older groups of users. The average cosine similarity became negative for all groups after 35-44. This behavior could mark a binary split between "young" and "old" users. Interestingly, we notice variation between the "young" and "old" groups. This trend also holds for the first scenario. However, we find a wider gap between the groups under 18 and 18-24, signaling that the 18-24 group does experience higher levels of age bias than other groups.

When testing the significance of the results for the BPR model, we found that the first scenario passed all significance tests. The second scenario also resulted in significance for the holistic significance tests. However, we found that not all results were significant when testing by group. The age group for 18-25 and over was found to not be significant, with a p-value of 0.016, when comparing differences in the cosine similarity between that of users and the age direction versus the previously calculated gender direction. All other groups had

significant differences in their cosine similarities with the age direction versus that with a random or gender direction. Since the first scenario resulted in statistical significance for all groups, we recommend moving forward with that bias direction instead of the binned age bias direction.

The DMF model did not result in the same significance levels, with both the first and second scenarios resulting in non-significance at the group level. Testing the first scenario resulted in non-significant tests for two user groups 50-55 and 56 and over, when comparing the age and gender bias direction. When testing against the random direction, the user groups representing ages between 45 to 56 and over were found not to be significantly different. For the second scenario, the binned age bias direction does not result in significantly different cosine similarities against the gender bias direction for the age group 50-55 and 56 and over. However, it does result in statistical significance across all user groups when tested against cosine similarities with the randomly generated bias direction.

One can leverage these statistical testing results to understand if the grouping strategies accurately reflect age bias within the latent space for every age group. Patterns in the testing results can signal when and how a different grouping strategy could be leveraged to achieve ultimate results. For example, with the DMF model, our results are insignificant for age groups over 45. After binning, only age groups over 50 have no significance. Based on this, one could reconfigure the binning to reflect these results by leveraging the groups under 18, 18-25, 35-55, and 56 and over. Another option could be to create groups under 18, 18-25, 35-45 and 45 and over. In the case of no patterns in significance, it may be best practice to leverage pairwise binary bias direction creation instead of attempting to find the optimal grouping for a holistic bias direction.

For the NCF MF embedding results, we find that our scenarios result in even less statistical significance, which is to be expected given the algorithmic differences. Results are insignificant against the randomly generated bias direction at the group level for age groups 35-55. Results are not significant at the holistic level when compared with the gender direction. For the second binned scenario, results are insignificant against the randomly generated direction for under 18 to 35 and 50-55. Additionally, the second scenario does not achieve significance for the group holistic test against the gender bias direction. Unlike the binary gender bias direction, NCF MLP embeddings result in more significant tests than the MF embeddings when results are compared against a randomly generated bias direction. For the first scenario, only the user group 56 and over is found not to be significantly different from the random direction. In the first scenario, age direction does not result in significance at a holistic level and all user group levels from the user gender direction. The second scenario results in all significant tests against the random bias direction and the holistic comparison against the gender bias direction. However, it does not showcase significance from the gender direction at the group level, with only the two age groups representing ages 25-50 showcasing significant tests.

Given these results, the only direction that passed significance tests allowing for further evaluation was the first scenario for the BPR model. Unlike previous work with binary bias directions, exploring multi-categorical bias directions is not the focus of this dissertation. Thus, we do not investigate non-significant bias directions for bias metrics in future sections. Future iterations of this work could explore this type of AAB direction in more depth, but that is out of scope for this work.

	BPR	DMF	NCF	
			MF	MLP
Under 18	0.2040	-0.1259	0.1452	-0.0101
18-24	0.2298	-0.1381	0.1731	-0.0225
25-34	0.0933	-0.0684	0.0401	-0.0262
35-44	-0.0304	0.0045	-0.0840	-0.0162
45-49	-0.0748	0.0333	-0.1281	-0.0085
50-55	-0.1028	0.0580	-0.1598	0.0010
Over 55	-0.1262	0.0726	-0.1742	0.0182

Table 6.6: Cosine similarity with the multi-group SVC bias direction for age created from the binned ages of under 18, 18-49, and over 50. One can see that there is visible opposing relationships between younger and older age groups.

			NCF	
	BPR	DMF	MF	MLP
Under 18	0.0784	-0.1245	0.1451	0.0254
18-24	-0.0570	-0.1175	0.0001	-0.0005
25-34	-0.0835	-0.0536	-0.0411	-0.0061
35-44	-0.0464	0.0098	0.0136	-0.0007
45-49	-0.0477	0.0382	0.0330	0.0006
50-55	-0.0563	0.0601	0.0243	-0.0015
Over 55	-0.0452	0.0793	0.0312	0.0082

Table 6.7: Cosine similarity with the multi-group SVC bias direction for age created from the unbinned ages originally reported in the data. One can see that unlike results for the binned age direction, only the DMF model resulted in noticeable behavior showcasing opposing relationships between younger and older age groups with the resulting bias direction.

Chapter 7

Implement: Significance

In this chapter, we will demonstrate how to identify significant attribute association bias concerning relationships between the attribute defining entity sets and test entity sets. We will showcase our methods for the two case studies: Spotify podcast recommendation and MovieLens movie recommendation. For Spotify podcast recommendation, we will focus on bias metrics. We do not showcase implementation of our other proposed methods for the Spotify case study since those implementations were out of scope for the original project. All methods will be demonstrated for the MovieLens movie recommendation case study.

7.1 Spotify Podcast Recommendation

In the previous chapter, we demonstrated the significance of the user gender bias direction between male and female podcast listeners. Now we will showcase results when calculating bias metrics to understand the relationship between user gender and two genres of podcasts: True Crime and Sports.

7.1.1 Bias Amplification Metrics

EAA, GEAA, and DEAA Our results using this set of metrics signaled that our test entity sets of true crime and sports podcast vectors showed significant association attribute bias with their respective user gender. We found that podcast embeddings trained with and without gender resulted in a significant DEAA score of 612.27 and 480.59, respectively. The calibration effect for with and without gender DEAA was 1.81 and 1.78. The normalization of the calibration effect showcases that the AAB remains highly significant when accounting for the EAA distributions.

EAA metrics successfully flagged a significant change in AAB levels when removing gender. However, significant levels of bias remained. When accounting for the separate GEAA for sports and true crime podcasts, we found that the final DEAA score could be contributed primarily to sports podcasts versus true crime podcasts. When trained with gender, sports podcasts GEAA was -521.34, which was reduced to -406.59 when trained without gender. This decrease of 22% was greater than the 18.6% decrease for the true crime podcast test metric. True crime GEAA was originally 90.93 when trained with gender and reduced to 73.98 after the mitigation. This discrepancy reflects that sports podcasts have significantly higher AAB and are heavily associated with the male attribute-defining entity set of vectors. This difference also highlights that this simple mitigation method does not equally address gender AAB across groups. Observing the different levels of EAA metrics allow a practitioner to pinpoint which group is more or less affected by the mitigation.

R-RIPA We find that R-RIPA also successfully relays AAB; results are in Table 7.1. R-RIPA results show fluctuations across the bias directions used,

indicating the importance of selecting a bias direction. In particular, our R-RIPA results using the PCA bias direction support our earlier suspicion that the bias direction did not accurately capture the user gender AAB direction. We see this because the R-RIPA significantly differs between the two podcast entity groups, but the R-RIPA maintains the same negative direction for both groups. All our R-RIPA results were statistically significant, with p ¡ 0.05. Our R-RIPA metrics were significantly higher than R-RIPA calculated via permutations across all podcasts.

When comparing the R-RIPA across bias directions, a couple of results stand out. First, after removing gender, the SVC R-RIPA for true crime podcasts increased, signifying an increase in AAB for true crime podcasts with female users. However, this result is not present for R-RIPA created with the bias directions CSVC-1, CSVC-2, and CD. This difference signals that user gender AAB may have a more nuanced relationship with individual female users that is not fully captured by centroid-based directions. Additionally, we find that true crime podcasts do not experience as significant of a decrease as sports podcasts for R-RIPA calculated with the bias directions CSVC-1, CSVC-2, and CD. This result could signify that removing user gender reduced AAB more heavily for sports podcasts with high levels of AAB as captured by a centroid-related direction.

Metric Comparison Unlike the EAA metrics, R-RIPA is at risk of more fluctuation in results depending on the bias direction selected for calculation. As a result, we recommend that practitioners compute R-RIPA only with bias directions that more accurately represent the attribute behavior in the latent space. For example, when leveraging bias directions other than that of SVC (trained on a randomly sampled user set), there is a significant increase in AAB signaled

by R-RIPA. This peculiarity could be seen as those bias directions over-reporting bias or SVC under-reporting bias. Additionally, R-RIPA computed with the SVC bias direction is at risk of becoming less accurate as the trained SVC becomes less accurate. It is essential to account for this possibility by implementing permutation testing to determine the significance of R-RIPA results if there is less confidence in the bias direction. In such cases, it may be more prudent to apply the EAA bias metrics instead.

	Bias Direction Used for Calculating R-RIPA									
	SVC		CSVC-1		CSVC-2		CD		PCA	
	Model Training Data									
	WG	NG	WG	NG	WG	NG	WG	NG	WG	NG
	Podcast Genre R-RIPA									
True Crime	0.132	0.182	0.118	0.115	0.137	0.131	0.138	0.129	-0.012	0.054
Sports	-0.142	-0.141	-0.239	-0.062	-0.234	-0.187	-0.234	-0.209	-0.061	-0.151
	R-RIPA Summary									
Differential	0.274	0.323	0.357	0.177	0.371	0.318	0.372	0.338	0.049	0.205
Effect	1.686	1.768	1.787	1.339	1.812	1.763	1.808	1.784	0.494	1.430

Table 7.1: R-RIPA results for podcast embeddings (trained with gender (WG) and without gender (NG)) for true crime or sports podcasts leveraging gender directions created from our bias direction methods. Negative and positive results signify male and female association, respectively.

7.2 MovieLens Movie Recommendation

After evaluating the significance of the bias directions, shown in the previous chapter, we flag genres for bias metric evaluation. We do this for the significant modeling scenarios BPR 1M, BPR 100K, and DMF 1M. Finally, we calculate and evaluate bias metrics for the two modeling scenarios with the most significant attribute association bias, BPR and DMF trained on the 1M MovieLens dataset.

7.2.1 Flagging Groups

For the analysis in this section, it is essential to note that negative EAA values reflect relationships with the female gender, and positive EAA values signal association with the male gender. We ran this analysis against all significant bias directions: BPR 100K SVC, BPR 1M SVC, BPR 1M Centroid, DMF 1M SVC, and DMF 1M Centroid. For these analyses, we set the target prediction value as the cosine similarity between the item and the significant direction (i.e., the individual direction R-RIPA). We find that algorithms showcase unique behaviors in how AAB relates to specific movie genres. We discuss both the final results and differences found below.

We do not report on NCF results due to the fact that algorithmically, the embeddings are NOT modeled into the same space until they are concatenated for final predictions. This holds true for the upcoming bias metric result section as well. We found that analyzing the entity embeddings separately did result in some levels of significance, but due to the embeddings being in separate spaces, these results are spurious and can be attributed to overall similarity between entity group behaviors. Leveraging these methods on entity embeddings in different latent spaces is not considered good practice and could lead to erroneous results.

While calculating the pairwise metrics, we found that genres on average showed some significant difference between each other in regards to their metric differences. Due to this, it may be prudent to test for significance by comparing the differences across the pairwise genres, thus allowing the practitioner to pinpoint which pairings are more or less biased towards different attributes. When looking for AAB, one can look for opposing biases between entity attributes or at biases of singular entity attributes. In order for biases to be opposing, the two calculated entity bias metrics must have opposing signs. In order to determine if there are specific groups that show more bias, one should test their test statistic against both the absolute value test statistic of all pairwise combinations and then against those with opposing bias. This allows one to observe if the bias is different across attributes and if it is more opposing.

Additionally, there may be scenarios where all categorical entity attributes exhibit significant levels of AAB based on permutation testing. In that case, one can deem the attribute to be strongly embedded into the latent space. Issues arise if certain categories exhibit significantly more bias, thus leading to certain stakeholders experiencing more representation harm in terms of serving stereotyped content.

If the categories are not mutually exclusive then one can look to the categorical analysis performed to compare groups of attributes together. For example, if romance, children's, and musical genres were found to be significant towards female users and action, western, and sci-fi to male users, one could calculate the metrics between those two groups of genres.

Given the overhead attributed with pairwise comparisons, we recommend exploring test entity attribute bias informed by qualitative studies and biases recommended for analysis by experts within the area. If that is not available,

one may wish to conduct the analysis framework described above. One benefit to conducting these proposed evaluation steps is that it may uncover lesser known stereotypes or behaviors within the unique recommendation modeling scenario.

Our analysis in these sections lead to the following key takeaways:

Decision Trees result in more accuracy The results across analysis types, linear regression and decision trees, remain relatively stable for our modeling scenarios. However, the decision trees showed higher levels of accuracy than our linear regression analysis. Given this, we look to the final decision tree results for calculating our AAB metrics. In particular, we find Action, Romance, Children's, Sci-Fi, Drama, War, Western, and Crime to all be significant predictors across the various modeling scenarios, thus we leverage those groupings for our analysis of each scenario.

Qualitative studies versus quantitative results Studying genre preference by gender has been studied via qualitative studies to understand the differences between perceived gender stereotypes and the actual differences in preference across gender. A previous study found that perceived stereotypes for female film genres were animation, comedy, drama, heimat, and romance. Alternatively, they found that male perceived genres included action, adventure, erotic, fantasy, history, horror, science fiction, thriller, war and western. Out of the seventeen genres studied, only two were classified as gender neutral: crime and mystery. When comparing with actual gender preferences, only drama and romance genres were found to be significantly preferred by women. Actual genre preferences for male participants in the study were action, adventure, erotic, fantasy, horror, mystery, science fiction, war, and western. [90]. These studies provide helpful

context for analyzing our results. For example, the qualitative study found that crime was gender neutral, but in our analysis it was one of the top predictors for male AAB scores. Beyond that difference, our quantitative findings were generally in line with the qualitative findings. In fact, the quantitative results reflected the perceived stereotypes more frequently than the actual stereotypes observed. Comparisons between results of our evaluation versus the qualitative studies can be found in Table 6.13. This table best showcases the lack of stereotyped attribute association bias in the NCF modeling scenario. It does not result in significant stereotyped AAB, which is due to how items and users are not embedded into the same space (as well as abstractions between final results and embeddings). Even though stereotyped AAB is not present, one can deduce the clustering of certain movie genres, such as the male versus female grouped genres for the NCF MF embeddings.

Modeling scenarios showcase unique AAB behaviors If one looks at the heat maps of the four modeling scenarios, one could observe that the correlations between the groups have a relationship with the strength of bias direction found in each scenario and how it relates to the various genre groups. For example, the strongest bias direction was found for BPR trained on the one million movielens dataset, we can see how the genres relate more strongly to the calculated cosine similarity with said SVC bias direction. The heatmap shows that one of the least correlated genres is between genres with contrasting gender stereotypes such as action and romance, or sci-fi and drama. The strongest negative correlation is between comedy and drama, but this is a well known difference in movie classifications, which is not necessarily linked to gender specific preferences.

Binary User Gender

In the following passages, we will cover the results when flagging groups for future evaluation of binary user gender attribute association bias. We cover the results for significant bias directions and provide insight into how NCF is not appropriate for future evaluation steps.

BPR 1M The linear SVC model achieves higher accuracy with a larger sample size. Additionally, we find the flagged groups to differ with more group features showing similar test statistics. For example, Sci-fi, War, and Western have test statistics between 12 and 14. While for female-associated features, the Musical group gains similar importance to Children's movies. However, Action and Romance remain the strongest features for determining a male versus female SVC individual R-RIPA score. The group of important attributes changes when observing the results from the decision tree, with Action, Romance, Sci-Fi, Drama, Crime, and War being the most important attributes for the decision tree with an accuracy of 0.89. The leaves of the decision tree regressor further confirm this behavior. However, it differs from the 100k sample, with Romance being the most important differentiator in the path of the decision tree. These differences in results highlight the importance of evaluating AAB when the training dataset changes.

We see a difference in behavior when comparing the analysis results for the Centroid versus SVC-based R-RIPA as the target prediction value. The accuracy of both the linear regression and decision tree increase significantly. The R2 value for predicting centroid R-RIPA is 0.60, while that for SVC R-RIPA is 0.425. The accuracy for the centroid R-RIPA decision tree is 0.97, while the SVC R-RIPA is 0.88. A change in feature importance also accompanies this difference in accuracy.

For the Centroid R-RIPA, we find Action, Drama, Sci-Fi, Romance, and Horror to be the most important group variables for the decision tree. For the SVC R-RIPA, we see that Action, Crime, Horror, Sci-Fi, and Western contribute to a more male prediction. Romance, Children's, Drama, and Musical contribute to a more female prediction in both models. We can see that group relationships differ as well with the heatmaps. For example, in the Centroid R-RIPA there is a stronger relationship between the Romance and Comedy feature than in the SVC R-RIPA analysis. The Centroid R-RIPA analysis also shows stronger positive relationships across more genre features.

BPR 100k The linear regression performance for the 100k dataset for BPR is much lower than its SVC 1M sample counterpart, at 0.233 vs. 0.40. Even though accuracy is not a goal of the exploratory analysis, this discrepancy raises concern that the linear regression may not accurately capture the feature relationships. This finding is further confirmed when looking at the decision tree performance, 0.77. Even though accuracy is low, we see similar results with Action and Sci-Fi, resulting in a more male prediction. Unlike the 1M sample, Animation is shown to be more important as a predictor. This result is accompanied by Horror, War, and Western being found to be insignificant. In general, the test statistics for the 100k linear regression variables are lower than that of the 1M sample analysis.

When comparing the analysis results of the decision tree, more differences arise. For example, Romance is no longer among the top five most important variables. Instead, the decision tree results mimic that of the Centroid R-RIPA decision tree for the 1M sample, with Drama being the second most important feature within the decision tree. The Children's genre feature also takes Sci-Fi's place as the third most important feature in the model.

DMF 1M Compared to the BPR models, Action and Sci-Fi remain essential. However, the Western genre shows a more significant relationship with the calculated individual SVC R-RIPA (as stereotypically male). In fact, in the linear regression results, it is found to have a higher test score than that of the Sci-Fi feature. We also see a unique shift in variable importance for the decision tree results, with the top five most important variables being Sci-Fi, Western, Action, War, and Romance. Unlike the BPR results, only one of these five is a more female predictor since Drama is no longer within the top five most important variables. Drama loses significance in two of the three classical methods for observing feature importance. Interestingly, when looking at the collinearity of the features, we find an increase in correlation between the groups and the output SVC R-RIPA, with strong collinearity between Sci-Fi, Action, and Adventure, and on the opposite spectrum, Musical, Animation and Children's movies. There is also a stronger collinearity between Romance and Comedy than in the BPR SVC results, which is only reflected in the BPR Centroid results.

The DMF Centroid R-RIPA analysis functions differently than the previously discussed scenarios. The decision tree accuracy is high at 0.97, but the top five genre variables are Action, Sci-fi, Drama, Comedy, and thriller. The collinearity heatmap also reflects more extreme relationships between variables, such as the highly positive collinearity between Children's and Animation, or the more negative relationship between Action and Drama. This difference showcases that the DMF model may exploit different latent relationships than the BPR algorithm. Even though the decision tree results show more differences, the linear regression results remain vaguely similar, with the top three male predictors being Action, Western, and Horror and more female scores being Romance, Musical, and Children's.

7.2.2 Bias Metrics

In this portion of the results overview, we will observe the SVC R-RIPA and EAA metric results across the two modeling scenarios, BPR and DMF trained on the 1M MovieLens dataset, between the flagged male genres of action, scifi, war, western and crime and the female genres romance, children's, and drama. We continued the audit with these two modeling scenarios given the significance of their bias directions and results when flagging for significant group relationships.

In total, this creates 15 comparisons for each modeling scenario. When filtering the movies for calculation, we needed to account for the fact that these genres were not mutually exclusive. We did this by grouping movies based on them being of a specified genre and not being categorized in any of the opposing gender genres. We tested the metrics for significance testing via permutation testing with a p-value of 0.00067 (leverages the Bonferroni correction for 15 tests). Additionally, we calculated the effect size of the genre comparison to understand the magnitude of the difference between the male and female genre being tested. Finally, we calculated the metrics for movies in any of the flagged male genre against movies in any of the flagged female genres. This comparison enables us to observe a broader interaction between male and female genres as well as account for movies that may be categorized in multiple genres.

For this analysis, we primarily look at calculated metrics for bias directions found to be significant. When we review the results we speak to both significance and effect size, we leverage effect size to compare significant pairings to determine where AAB for user gender is the strongest.

The R-RIPA scores for BPR and DMF show that there are significant biases between flagged female and male genres with female and male users. In BPR 1M,

we find that the pairings with the strongest effect size is that between war and romance, 2.2911, and western and romance, 2.2589. This result differs from the flagged groups highlighted for BPR, showcasing that it is important to account for stereotypes that may not be flagged by quantitative analysis, but qualitative analysis such as user research studies as well. It is important to note that these results were calculated for movies that are not classified as any of the opposing gender's categories. When this restraint is lifted, we find that results do change, pointing to the importance of genre interaction within the latent space. Future work for in depth analysis of this phenomenon would require creation of bias directions or metrics for multi-category scenarios, which is currently out of scope for this work.

We do not report on both EAA and centroid R-RIPA effect because the results are redundant. This is due to the both metrics capturing similar relationships, except they are calculated in different ways, thus the base distribution comparisons are more or less of the same magnitude.

Centroid R-RIPA

Out of the two modeling scenarios, we find BPR to result in higher levels of Centroid R-RIPA as shown by the greater differential between male and female stereotyped genres. Action and Romance have the highest Centroid R-RIPA for the BPR model with 0.330 and -0.3225, respectively. In the DMF model, Sci-Fi shows the most AAB towards male users than Action. Overall, male stereotyped genres result in higher Centroid R-RIPA across both modeling scenarios.

	BPR			DMF		
	R-RIPA		GEAA	R-RIPA		GEAA
	Centroid	SVC		Centroid	SVC	
Action	0.3300	0.1677	31.7846	0.1606	0.0488	10.2351
Sci-Fi	0.3222	0.1601	20.3402	0.1798	0.0418	7.5111
War	0.2732	0.2421	4.0550	0.1078	0.0553	1.0584
Western	0.2107	0.2164	2.9001	0.0983	0.0738	0.8951
Crime	0.1763	0.0928	5.2797	0.0605	-0.0067	1.1989
Romance	-0.3225	-0.1879	-34.3704	-0.1495	-0.1613	-10.5441
Drama	-0.1703	-0.0712	-56.8799	-0.0585	-0.0994	-12.9407
Children's	-0.1168	-0.0867	-6.9995	-0.0753	-0.0617	-2.9834

Table 7.2: This table reflects results for R-RIPA when calculated with the Centroid and SVC user gender bias direction and the GEAA for the MovieLens BPR and DMF recommendations created from the 1M sample dataset.

SVC R-RIPA

SVC R-RIPA results are completely different from the Centroid R-RIPA. We find the scale of the results to be smaller, which is reflected in the metric effect tables as well. Similar to Centroid R-RIPA, BPR shows higher levels of AAB across all genres. We find a lessening in AAB for Action and Romance. Additionally, we find that for DMF Crime no longer results in significant AAB. According to SVC R-RIPA, War and Western movies are the most male stereotyped movie genres, not action and Sci-Fi. This difference could be due to the linear SVC model resulting in more nuanced results which is lost when leveraging a more aggregative metric such as Centroid R-RIPS or EAA.

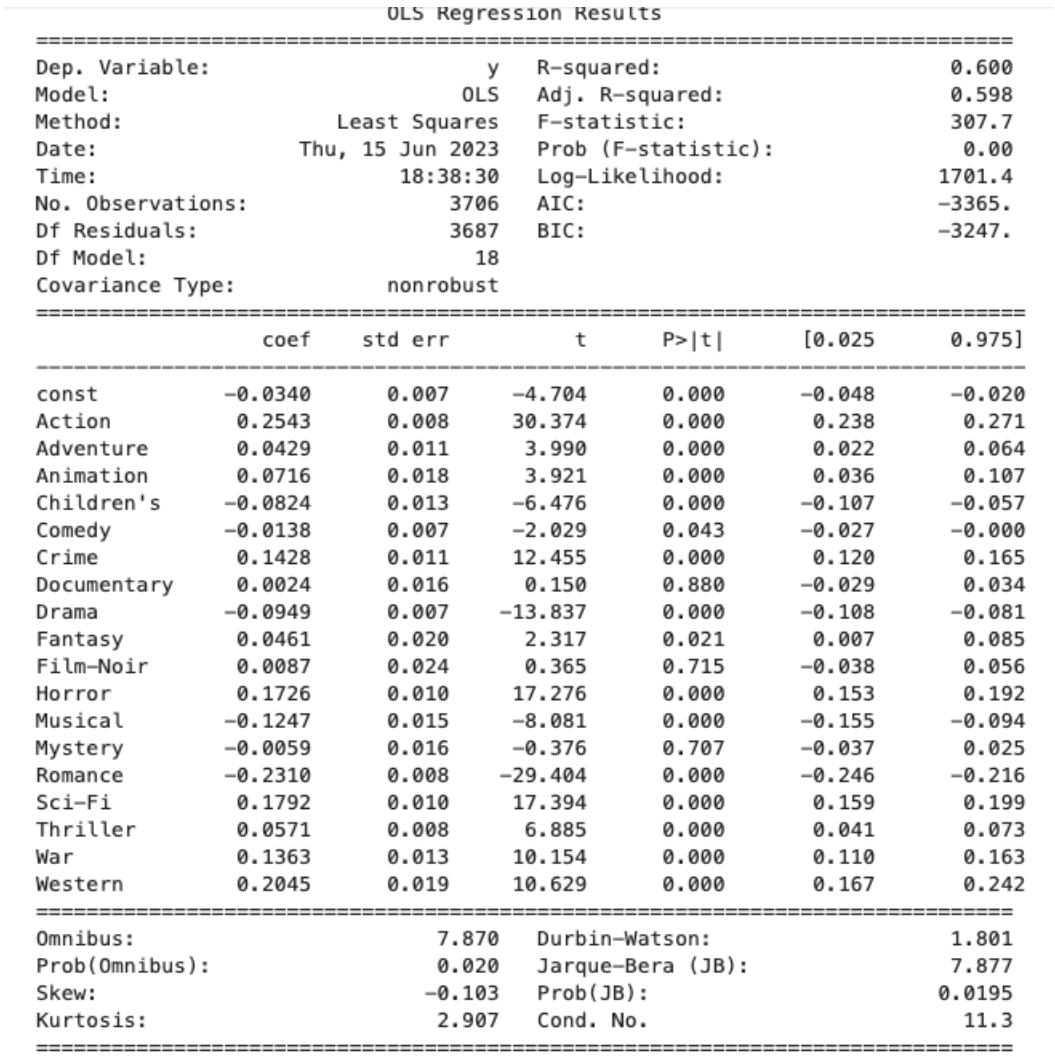
EAA

We find EAA metrics to showcase slightly different results from that of R-RIPA, particularly SVC R-RIPA. Action and Sci-Fi result in the highest EAA across

		BPR		DMF	
		SVC	Centroid	SVC	Centroid
Action	Romance	1.6738	1.8119	1.3484	1.5346
	Children's	1.5704	1.7549	1.6088	1.5939
	Drama	1.4386	1.8409	1.0751	1.6384
Crime	Romance	1.6675	1.8619	1.0254	1.2829
	Children's	1.3371	1.4488	1.1996	1.3255
	Drama	1.1345	1.6752	0.5551	1.1798
Sci-Fi	Romance	1.6953	1.8469	1.2746	1.5663
	Children's	1.5373	1.7103	1.4882	1.5944
	Drama	1.4554	1.9483	1.0083	1.8073
War	Romance	2.2911	2.2912	1.3543	1.5331
	Children's	1.9599	1.8351	1.7253	1.6654
	Drama	2.0553	2.1624	1.1467	1.6286
Western	Romance	2.2589	2.2450	1.4400	1.5164
	Children's	1.9175	1.9175	1.7886	1.7205
	Drama	1.9206	1.9312	1.3101	1.5682

Table 7.3: This table reflects results for the R-RIPA effect when calculated with the Centroid and SVC user gender bias direction for the MovieLens BPR and DMF recommendations created from the 1M sample dataset.

both BPR and DMF models. The difference we find is the highest EAA towards female users is Drama. As previously stated, EAA reflects the aggregation of individual entity-to-entity cosine similarities. This difference between EAA and R-RIPA may signal that a higher number of female users show association with the Drama genre. We also see a lessening in reflected bias for War and Western movies. This paired with R-RIPA results could signal that a smaller group of male users could be driving those results.



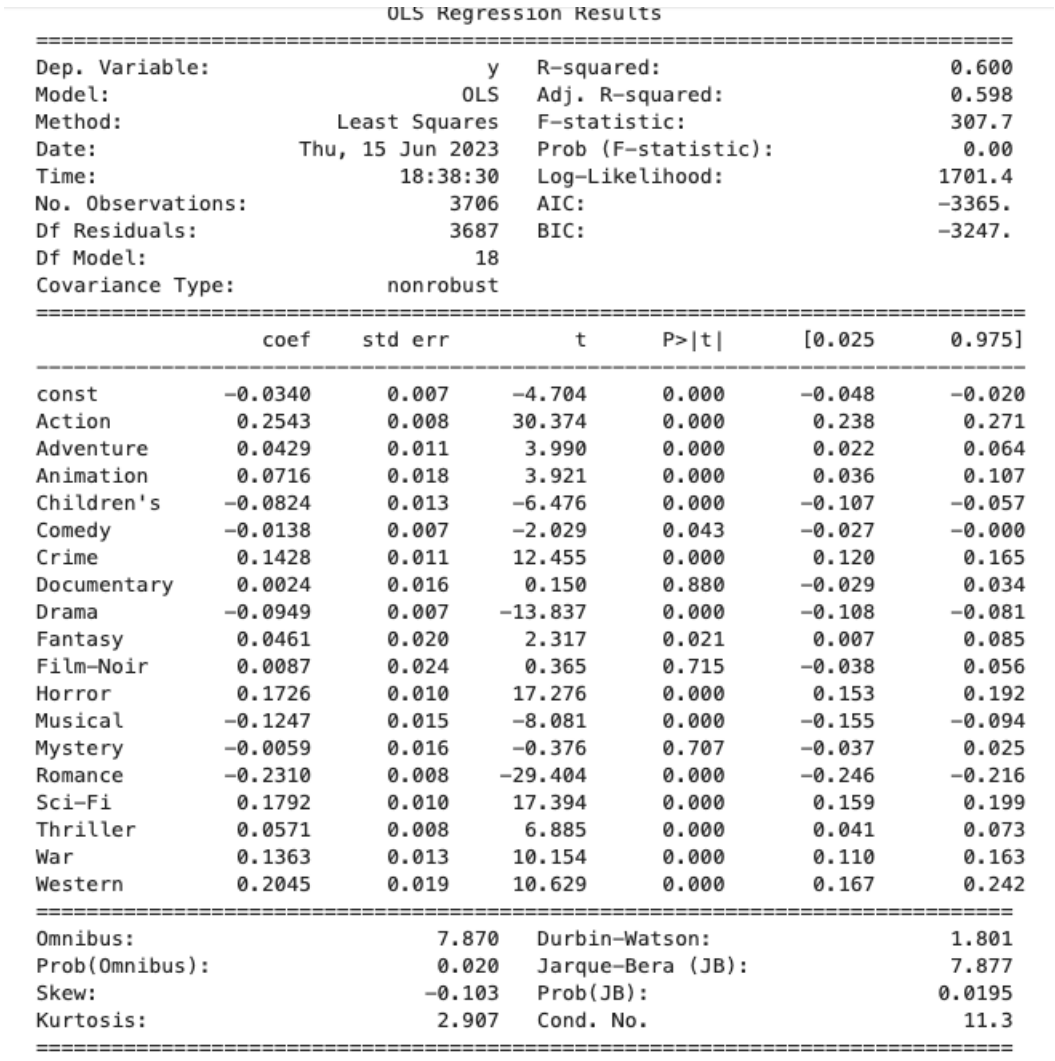
OLS Regression Results

Dep. Variable:	y	R-squared:	0.600
Model:	OLS	Adj. R-squared:	0.598
Method:	Least Squares	F-statistic:	307.7
Date:	Thu, 15 Jun 2023	Prob (F-statistic):	0.00
Time:	18:38:30	Log-Likelihood:	1701.4
No. Observations:	3706	AIC:	-3365.
Df Residuals:	3687	BIC:	-3247.
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0340	0.007	-4.704	0.000	-0.048	-0.020
Action	0.2543	0.008	30.374	0.000	0.238	0.271
Adventure	0.0429	0.011	3.990	0.000	0.022	0.064
Animation	0.0716	0.018	3.921	0.000	0.036	0.107
Children's	-0.0824	0.013	-6.476	0.000	-0.107	-0.057
Comedy	-0.0138	0.007	-2.029	0.043	-0.027	-0.000
Crime	0.1428	0.011	12.455	0.000	0.120	0.165
Documentary	0.0024	0.016	0.150	0.880	-0.029	0.034
Drama	-0.0949	0.007	-13.837	0.000	-0.108	-0.081
Fantasy	0.0461	0.020	2.317	0.021	0.007	0.085
Film-Noir	0.0087	0.024	0.365	0.715	-0.038	0.056
Horror	0.1726	0.010	17.276	0.000	0.153	0.192
Musical	-0.1247	0.015	-8.081	0.000	-0.155	-0.094
Mystery	-0.0059	0.016	-0.376	0.707	-0.037	0.025
Romance	-0.2310	0.008	-29.404	0.000	-0.246	-0.216
Sci-Fi	0.1792	0.010	17.394	0.000	0.159	0.199
Thriller	0.0571	0.008	6.885	0.000	0.041	0.073
War	0.1363	0.013	10.154	0.000	0.110	0.163
Western	0.2045	0.019	10.629	0.000	0.167	0.242

Omnibus:	7.870	Durbin-Watson:	1.801
Prob(Omnibus):	0.020	Jarque-Bera (JB):	7.877
Skew:	-0.103	Prob(JB):	0.0195
Kurtosis:	2.907	Cond. No.	11.3

Figure 7.1: OLS regression results when predicting Centroid R-RIPA leveraging genres as features for the MovieLens BPR model trained on the 1M sample dataset.



OLS Regression Results

Dep. Variable:	y	R-squared:	0.600
Model:	OLS	Adj. R-squared:	0.598
Method:	Least Squares	F-statistic:	307.7
Date:	Thu, 15 Jun 2023	Prob (F-statistic):	0.00
Time:	18:38:30	Log-Likelihood:	1701.4
No. Observations:	3706	AIC:	-3365.
Df Residuals:	3687	BIC:	-3247.
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0340	0.007	-4.704	0.000	-0.048	-0.020
Action	0.2543	0.008	30.374	0.000	0.238	0.271
Adventure	0.0429	0.011	3.990	0.000	0.022	0.064
Animation	0.0716	0.018	3.921	0.000	0.036	0.107
Children's	-0.0824	0.013	-6.476	0.000	-0.107	-0.057
Comedy	-0.0138	0.007	-2.029	0.043	-0.027	-0.000
Crime	0.1428	0.011	12.455	0.000	0.120	0.165
Documentary	0.0024	0.016	0.150	0.880	-0.029	0.034
Drama	-0.0949	0.007	-13.837	0.000	-0.108	-0.081
Fantasy	0.0461	0.020	2.317	0.021	0.007	0.085
Film-Noir	0.0087	0.024	0.365	0.715	-0.038	0.056
Horror	0.1726	0.010	17.276	0.000	0.153	0.192
Musical	-0.1247	0.015	-8.081	0.000	-0.155	-0.094
Mystery	-0.0059	0.016	-0.376	0.707	-0.037	0.025
Romance	-0.2310	0.008	-29.404	0.000	-0.246	-0.216
Sci-Fi	0.1792	0.010	17.394	0.000	0.159	0.199
Thriller	0.0571	0.008	6.885	0.000	0.041	0.073
War	0.1363	0.013	10.154	0.000	0.110	0.163
Western	0.2045	0.019	10.629	0.000	0.167	0.242

Omnibus:	7.870	Durbin-Watson:	1.801
Prob(Omnibus):	0.020	Jarque-Bera (JB):	7.877
Skew:	-0.103	Prob(JB):	0.0195
Kurtosis:	2.907	Cond. No.	11.3

Figure 7.2: OLS regression results when predicting Centroid R-RIPA leveraging genres as features for the MovieLens BPR model trained on the 1M sample dataset.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.425
Model:	OLS	Adj. R-squared:	0.422
Method:	Least Squares	F-statistic:	151.1
Date:	Thu, 15 Jun 2023	Prob (F-statistic):	0.00
Time:	18:38:15	Log-Likelihood:	2699.2
No. Observations:	3706	AIC:	-5360.
Df Residuals:	3687	BIC:	-5242.
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0148	0.006	2.679	0.007	0.004	0.026
Action	0.1094	0.006	17.104	0.000	0.097	0.122
Adventure	-0.0018	0.008	-0.215	0.830	-0.018	0.014
Animation	0.0429	0.014	3.079	0.002	0.016	0.070
Children's	-0.0811	0.010	-8.344	0.000	-0.100	-0.062
Comedy	-0.0041	0.005	-0.788	0.431	-0.014	0.006
Crime	0.0784	0.009	8.951	0.000	0.061	0.096
Documentary	-0.0038	0.012	-0.310	0.757	-0.028	0.020
Drama	-0.0536	0.005	-10.217	0.000	-0.064	-0.043
Fantasy	-0.0350	0.015	-2.302	0.021	-0.065	-0.005
Film-Noir	0.0155	0.018	0.846	0.398	-0.020	0.051
Horror	0.0150	0.008	1.968	0.049	5.71e-05	0.030
Musical	-0.0927	0.012	-7.861	0.000	-0.116	-0.070
Mystery	-0.0325	0.012	-2.719	0.007	-0.056	-0.009
Romance	-0.1679	0.006	-27.976	0.000	-0.180	-0.156
Sci-Fi	0.0978	0.008	12.425	0.000	0.082	0.113
Thriller	-0.0139	0.006	-2.202	0.028	-0.026	-0.002
War	0.1422	0.010	13.871	0.000	0.122	0.162
Western	0.1804	0.015	12.271	0.000	0.152	0.209

Omnibus:	3.432	Durbin-Watson:	1.810
Prob(Omnibus):	0.180	Jarque-Bera (JB):	3.478
Skew:	-0.069	Prob(JB):	0.176
Kurtosis:	2.941	Cond. No.	11.3

Figure 7.3: OLS regression results when predicting SVC R-RIPA leveraging genres as features for the MovieLens BPR model trained on the 1M sample dataset.

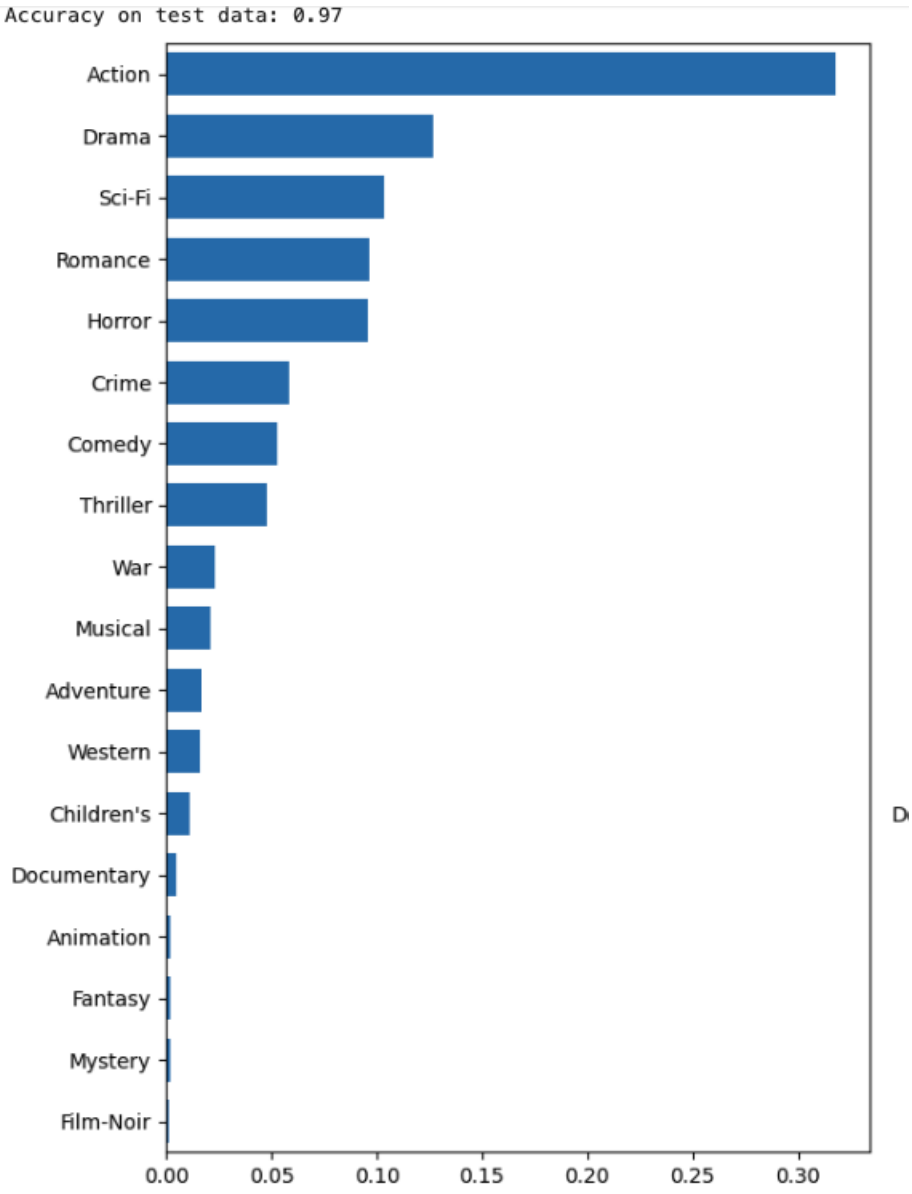


Figure 7.4: Tree variable importance when predicting Centroid R-RIPA leveraging genres as features for the MovieLens BPR model trained on the 1M sample dataset.

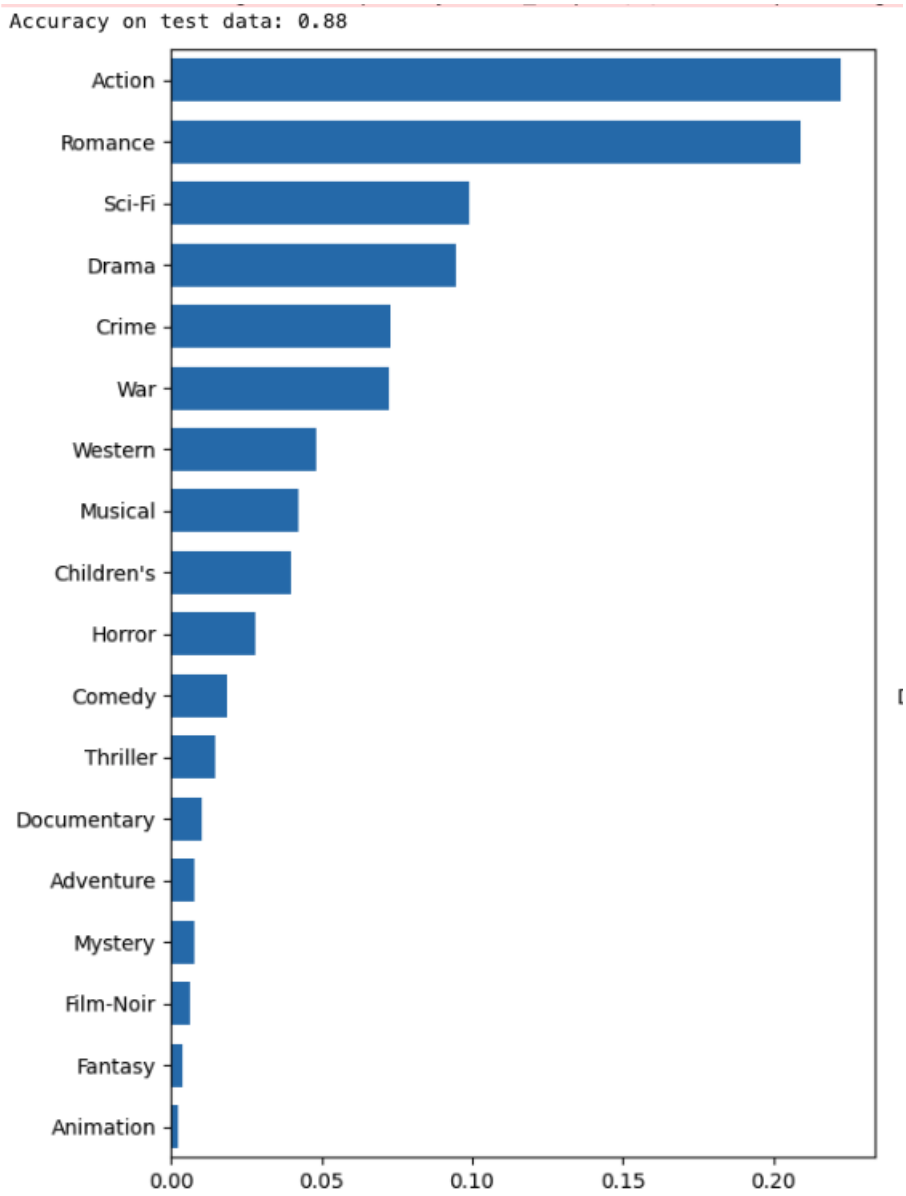


Figure 7.5: Tree variable importance when predicting SVC R-RIPA leveraging genres as features for the MovieLens BPR model trained on the 1M sample dataset.

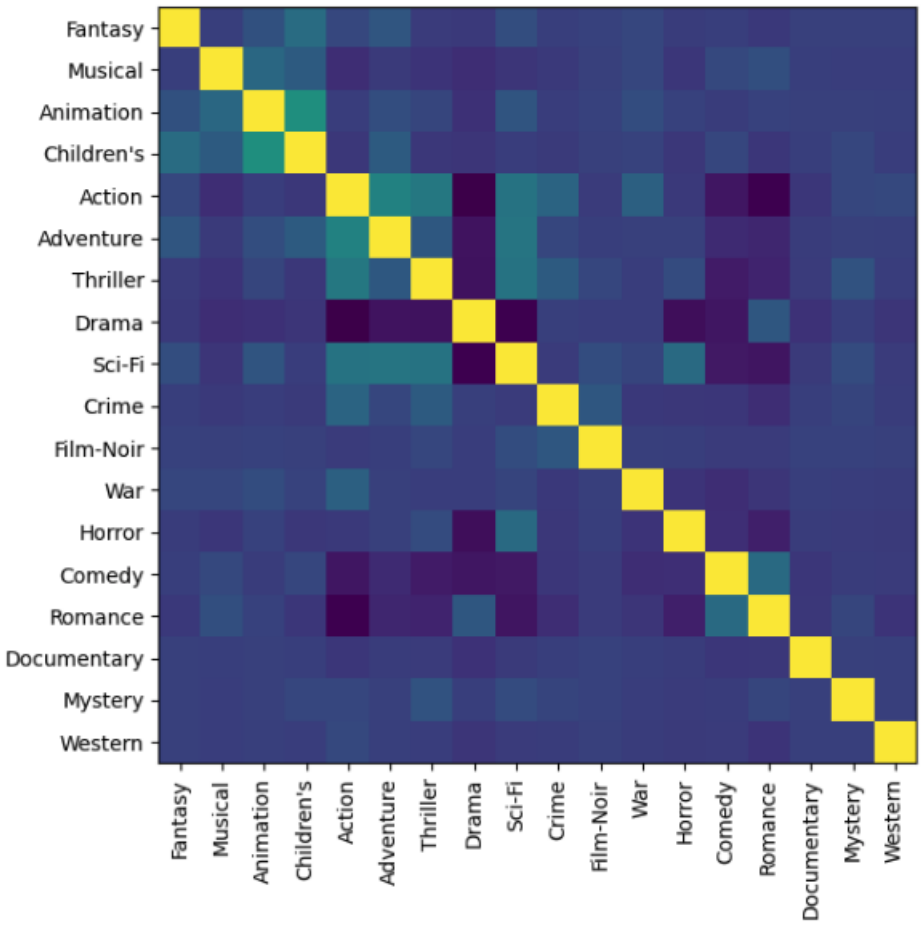


Figure 7.6: Collinearity heatmap when predicting Centroid R-RIPA leveraging genres as features for the MovieLens BPR model trained on the 1M sample dataset.

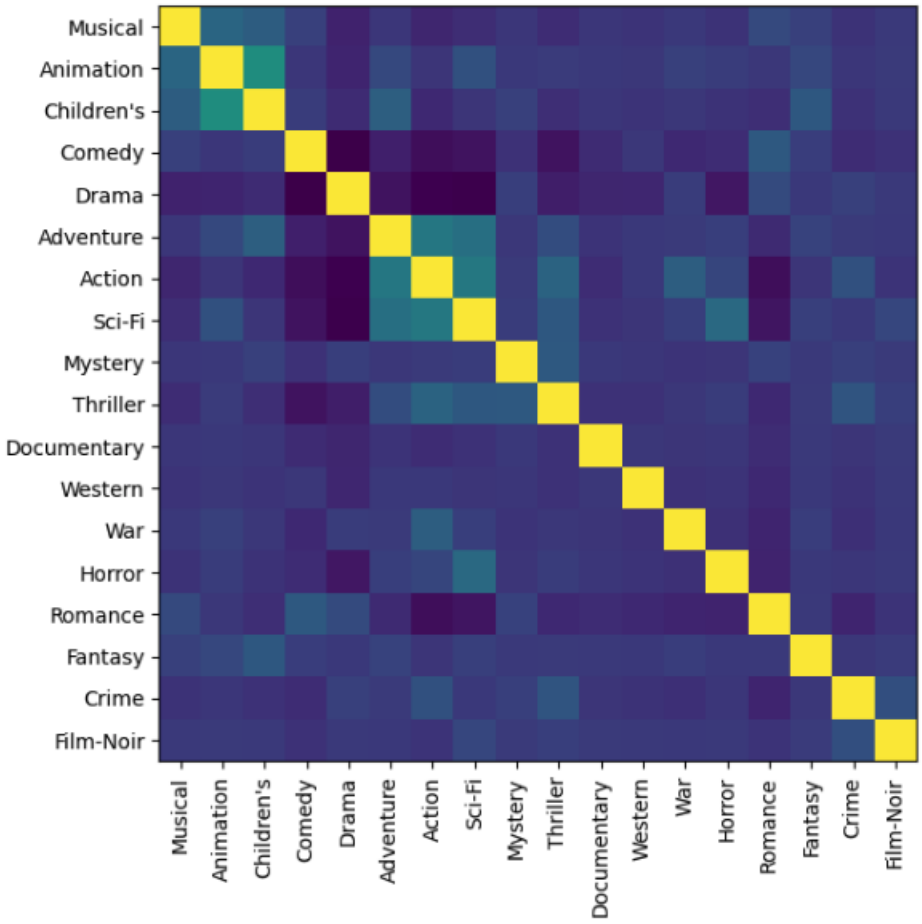


Figure 7.7: Collinearity heatmap when predicting SVC R-RIPA leveraging genres as features for the MovieLens BPR model trained on the 1M sample dataset.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.233
Model:	OLS	Adj. R-squared:	0.225
Method:	Least Squares	F-statistic:	28.05
Date:	Thu, 15 Jun 2023	Prob (F-statistic):	9.69e-83
Time:	19:10:57	Log-Likelihood:	1686.5
No. Observations:	1682	AIC:	-3335.
Df Residuals:	1663	BIC:	-3232.
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0120	0.006	-1.945	0.052	-0.024	0.000
action	0.0747	0.007	10.454	0.000	0.061	0.089
adventure	2.845e-05	0.009	0.003	0.998	-0.018	0.018
animation	0.0619	0.016	3.771	0.000	0.030	0.094
child	-0.0663	0.011	-6.283	0.000	-0.087	-0.046
comedy	0.0170	0.006	2.867	0.004	0.005	0.029
crime	0.0363	0.009	3.961	0.000	0.018	0.054
documentary	-0.0212	0.014	-1.524	0.128	-0.048	0.006
drama	-0.0289	0.006	-4.941	0.000	-0.040	-0.017
fantasy	0.0047	0.020	0.229	0.819	-0.035	0.045
noir	0.0247	0.019	1.281	0.200	-0.013	0.063
horror	0.0302	0.010	2.934	0.003	0.010	0.050
musical	0.0052	0.013	0.398	0.690	-0.020	0.031
mystery	0.0071	0.012	0.590	0.555	-0.017	0.031
romance	-0.0410	0.006	-6.463	0.000	-0.053	-0.029
scifi	0.0526	0.010	5.299	0.000	0.033	0.072
thriller	0.0043	0.007	0.615	0.539	-0.009	0.018
war	0.0281	0.011	2.558	0.011	0.007	0.050
western	0.0513	0.018	2.908	0.004	0.017	0.086

Omnibus:	12.012	Durbin-Watson:	1.786
Prob(Omnibus):	0.002	Jarque-Bera (JB):	16.936
Skew:	0.034	Prob(JB):	0.000210
Kurtosis:	3.487	Cond. No.	11.3

Figure 7.8: Tree variable importance when predicting SVC R-RIPA leveraging genres as features for the MovieLens BPR model trained on the 100k sample dataset.

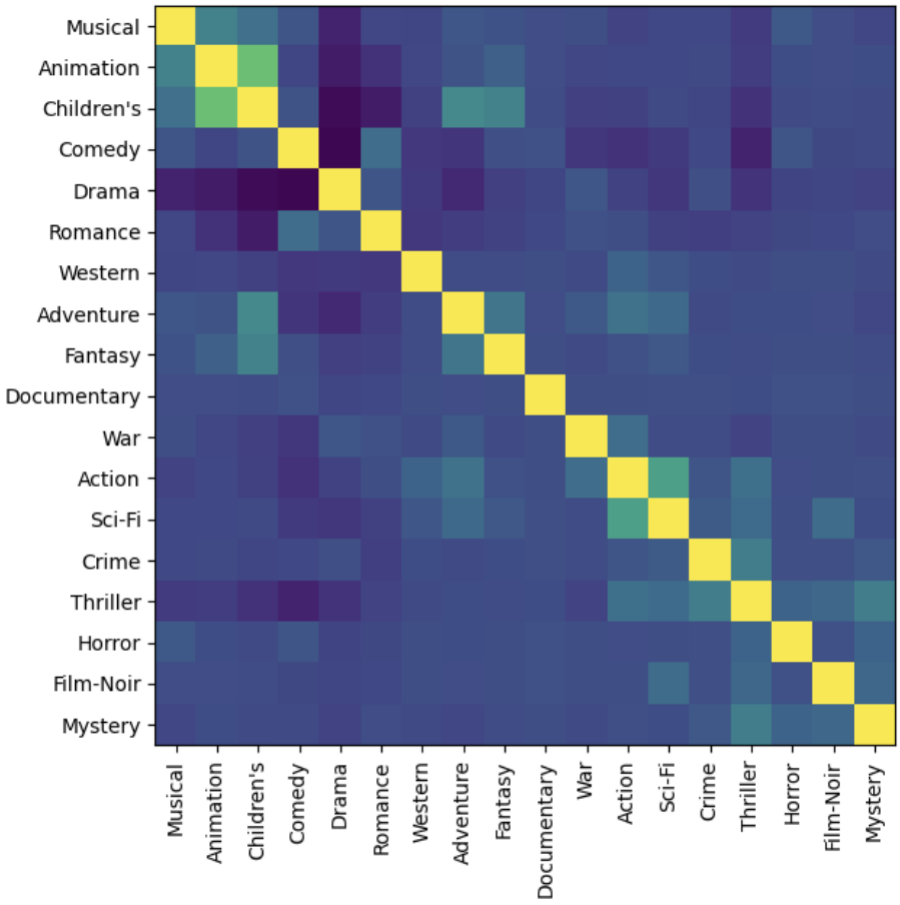


Figure 7.9: Collinearity heatmap when predicting SVC R-RIPA leveraging genres as features for the MovieLens DMF model trained on the 1M sample dataset.

Chapter 8

Implement: Amplification

In this final implementation chapter, we present our methods concerning the last step in the ESA framework: Amplification. We conduct this step for both case studies, but similar to previous chapters, our implementation for the Spotify case study is more limited than that of the MovieLens case study. We explore classification scenarios for both case studies. Feedback loop simulation is only explored for the MovieLens movie recommendation case study. For the feedback loop simulation, given the stable bias levels for each recommendation algorithm, we simulated feedback loops and examined embeddings for BPR and DMF trained on the 1M MovieLens dataset. By doing so, we can observe if specific algorithms are at a higher risk for amplifying bias and creating feedback loops for stereotypical recommendations over time.

8.1 Feedback Loops in Movie Recommendation

As previously noted, we evaluate the behavior of AAB over time with simulations of how recommendations would change if AAB is left unchecked. We do so by

simulating recommendations from BPR and DMF trained on the 1M MovieLens dataset. We run simulations five times and each training dataset is created by introducing previously recommended items into the new training datasets based on an acceptance probability function of (rewrite in math terms) e to the negative alpha multiplied by the rank.

8.1.1 Deep Matrix Factorization

The most notable difference when comparing bias directions between the original and final simulation entity vectors is the change in significance for the centroid bias direction. Originally the centroid bias direction was not found to be significant due to the lack of significance between male vectors and the centroid direction versus a randomly generated direction. However, after five simulation iterations, the average cosine similarity with the centroid direction increases significantly resulting in a more significant centroid bias direction. This is also demonstrated by the more equal distribution of cosine similarity between male and female users and the significant centroid bias direction resulting from the simulations. Even though, we find that average cosine similarity for female users and the bias directions become less pronounced, we see a drastic increase in average cosine similarity for male users and the centroid bias direction. This results in a wider distance between male and female users according to the bias direction.

We observe a similar change in sign for the male average cosine similarity with the SVC bias direction, resulting in male users having an opposite direction in comparison to female users according to the SVC bias direction. We found that male cosine similarities with the male direction of the vector increased significantly for both the SVC and centroid bias direction. Alternatively, female cosine

		Male Avg CS		Female Avg CS		Difference	
		1	5	1	5	1	5
BPR	SVC	0.0638	0.0676	-0.1178	-0.1037	0.1816	0.1713
	Centroid	0.1145	0.1209	-0.1552	-0.1557	0.1687	0.2766
DMF	SVC	-0.0392	0.0274	-0.1525	-0.0902	0.1133	0.1176
	Centroid	0.0070	0.0997	-0.1714	-0.0937	0.1784	0.1934

Table 8.1: This table showcases how cosine similarity between male and female users with the centroid and SVC bias direction change over simulations. Results are shown for the BPR and DMF model trained on the 1M MovieLens sample dataset.

similarities with both bias directions decreased significantly (meaning they became less similar to the female direction of the vector).

When comparing the first iteration to final iteration AAB metrics one can observe that stereotypically male genres experienced higher levels of bias after the feedback simulations. The genres Action, Sci-Fi, and War all experienced higher bias metrics for the final iteration. Between the different bias metrics, SVC R-RIPA resulted in the most significant changes across genres. Additionally, these changes were more drastic than those seen in Centroid R-RIPA and GEAA, which could be the result of SVC R-RIPA's capability to capture more nuances in behavior.

We can see that Centroid R-RIPA and GEAA both show Action and Drama experiencing significant results. GEAA also shows a significant increase in AAB for Children's movies towards female users. For stereotypically male genres, GEAA resulted in the highest delta between iterations. Unlike Centroid R-RIPA, GEAA reflects the aggregated difference of average cosine similarities between the individual entity vectors and each of the attribute entity vectors. This difference results in GEAA being more susceptible to highly biased outliers, which given results, may occur more often for male users and genres than that for female

users and genres.

The reduced robustness to outliers of GEAA is further demonstrated by the results of testing changes in the individual EAA bias metric distributions. These tests show that not every large GEAA delta is accompanied by a significant change in the EAA distribution. For example, GEAA deltas for War and Crime movies are the highest (0.2358 and 0.2773), but are not accompanied by significant test results. In these cases, one can assume that the increase or decrease is driven by outliers, not a population shift of bias.

The distinct increase for all stereotypically male genres point to the DMF algorithm reinforcing and strengthening the latent relationship between Action, Sci-Fi, and War movie vectors and male user vectors. It is important to note the imbalance within the original dataset between male and female users could be one of the underlying drivers in the higher deltas of bias for male genres due to the simulation creating more male interaction user instances with each simulation iteration.

8.1.2 Bayesian Personalized Ranking

Unlike the DMF algorithm, iterations of BPR recommendations do not result in a strong shift in the bias direction. The first and final iterations of bias directions remain significant, but do not result in sign changes like that of the DMF male average cosine similarity with the SVC bias direction. This lack of change could signal that the BPR algorithm is more resistant to reinforcing bias over time, resulting in potentially harmful feedback loops. We can see a slight increase in the average male and female cosine similarity with the centroid vector, but after testing the distributions, we found the shift to not be significant. When testing

the individual entity distributions and accounting for the Bonferroni correction, none of the changes in the cosine similarity distributions for both user genders were found to be significant. It is important to note, that unlike DMF, the original difference in cosine similarity between male and female users demonstrates opposite directions between genders and the calculated bias directions

Similar to significance results for the bias directions, individual bias metric significance tests resulted in finding that increase or decreases in the aggregate bias metrics were not accompanied by significant changes in the individual bias metric distributions. This result signals that these changes in aggregate metrics were the result of specific entities experience a higher or lower than usual change in their AAB with user gender.

Even though no significant changers occurred, it is interesting that the majority of genres experienced an increase in metrics. Action, Sci-Fi, and War SVC R-RIPA were the only genres to show signs of a decrease in bias. Since the metrics Centroid R-RIPA and GEAA both resulted in positive increases without significant results, one can hypothesize that certain outliers experienced higher levels of bias in comparison to the majority of users. This behavior could signal certain users being more at risk for reinforcing bias than others, which can have particularly harmful effects according to the content that is being served to the user.

8.2 Classification for Reinforcing Bias

In the following sections, we will showcase how classification can be leveraged to evaluate AAB and systematic bias in recommendation embeddings. We leverage three classification scenarios for evaluating AAB in Spotify podcast and Movie-

	BPR			DMF		
	1	5	Δ	1	5	Δ
			SVC R-RIPA			
Action	0.1677	0.1582	-0.0566	0.0489	0.0589	0.2045**
Sci-Fi	0.1602	0.1526	-0.0474	0.0418	0.0481	0.1507**
War	0.2421	0.2412	0.0037	0.0553	0.1023	0.8499**
Western	0.2164	0.2078	0.0397	0.0738	0.1195	0.6192**
Crime	0.0928	0.0902	-0.0280	-0.0067	0.0049	1.7313
Romance	-0.1879	-0.1902	0.0122	-0.1613	-0.1412	-0.1246
Drama	-0.0713	-0.0737	0.0336	-0.0617	-0.0743	0.2042**
Children's	-0.0867	-0.1003	0.1569	-0.0994	-0.0696	-0.2998**
			Centroid R-RIPA			
Action	0.3300	0.3545	0.0742	0.1606	0.1765	0.0990**
Sci-Fi	0.3222	0.3517	0.0916	0.1798	0.1984	0.1034
War	0.2732	0.2782	0.0179	0.1078	0.1229	0.1401
Western	0.2108	0.2220	0.0531	0.0983	0.0949	0.0346
Crime	0.1763	0.1914	0.0856	0.0605	0.0711	0.1752
Romance	-0.3226	-0.3399	0.0536	-0.1496	-0.1370	-0.0842
Drama	-0.1703	-0.1856	0.0898	-0.0585	-0.0474	-0.1897**
Children's	-0.1168	-0.1403	0.2012	-0.0753	-0.0851	0.1301
			GEAA			
Action	31.78	34.99	0.1010	10.23	12.18	0.1906**
Sci-Fi	20.34	22.76	0.1189	7.51	8.98	0.1957
War	4.06	4.23	0.0419	1.06	1.31	0.2358
Western	2.90	3.13	0.0793	0.89	0.94	0.0561
Crime	5.28	5.88	0.1136	1.19	1.52	0.2773
Romance	-34.37	-37.13	0.0803	-10.54	-10.47	-0.0066
Drama	-56.88	-63.55	0.1172	-12.94	-11.35	-0.1228**
Children's	-6.99	-8.61	0.2317	-2.98	-3.65	0.2248**

Table 8.2: This table showcases how bias metrics change over simulations. Significant changes are marked with **. Results are shown for the BPR and DMF model trained on the 1M MovieLens sample dataset.

Lens movie recommendation embeddings. The first scenario targets understanding how embeddings may relay gendered engagement with items by evaluating if item embeddings are accurately classfied as male or female in comparison to their gendered engagement patterns. The second scenario looks at gender bias by genre and if more stereotypically gendered genres also result in stereotyped classification results. The third, and final, scenario observes if user engagement history can be used to predict the gender of the user. This final scenario is particularly important due to privacy concerns if gender can be determined from engagement history.

For the Spotify Podcast Recommendation case study, we observe changes in classification results for when gender is used and not used during model training. When analyzing the MovieLens case study, we evaluate changes in results between the original model and the final fifth simulated recommendation embeddings. Similar to the previous section on feedback loops, we only observe changes for the BPR and DMF model.

8.2.1 Spotify Podcast Recommendation

The classification scenarios we designed allowed us to observe if podcast embeddings used as downstream features resulted in either accurate predictions of user gender engagement or stereotyped predictions of podcasts labeled for our entity test sets. For each scenario, we evaluated results for podcasts trained with and without user gender as a feature to understand implicit user gender bias in the latent space and how explicit use of the feature amplifies said bias. We used the same SVC classification models trained on user vectors to create gender directions for our analysis: SVC, CSVC-1, and CSVC-2.

Gendered Podcast Listening

We analyzed whether these predictions aligned with actual podcast listenership gender percentages. We did this by observing how our SVC models labeled podcasts as "male" and "female". We compared these predictions against the podcasts' male and female listenership percentage. In 8.3, we see the pattern that as podcasts have increasing percentages of male or female listenership, the podcasts are more likely to be classified as "male" or "female" podcasts. For example, with the SVC model trained on user embeddings with user gender, we see that when podcasts are in the 50% decile, they are classified as "female" 70.8% of the time, but when female listenership grows to over 70%, podcasts are labeled as "female" over 95% of the time. This classification scenario allows us to see that as engagement becomes more gendered, the podcast entity embeddings become more associated with a specific gender as well.

Interestingly, predictions correlating with female podcast listening became more accurate when the model was not trained with gender. However, this result did not hold for male podcast listening. When the model was trained without gender, the predictions became significantly less accurate when labeling a podcast with higher male engagement as male. Given this change in result, we hypothesized that the semantic embedding of user gender might not precisely represent the female and male binary relationship for podcast vectors but that of male and not male. Understanding how this relationship is embedded into the space would require more in-depth testing with non-binary data, which is out of the scope of this paper but could be an interesting development to explore in future research.

We found that this classification scenario showcased how podcast entity vectors can capture user gender AAB based on the increase in accuracy in predictions

as the percentage of listener gender rose. Additionally, the results showed that podcast embeddings associated with male listening experienced a sharper increase in accuracy as the male listener percentage increased. This finding is helpful during evaluation because it flags a difference in behavior within the latent space for podcast embeddings more related to stereotypical male listening.

% Decile of Listeners by Gender		Bias Direction Model											
		SVC				CSVC-1				CSVC-2			
		Model Training Data											
		WG		NG		WG		NG		WG		NG	
		Predicted Label for Podcast											
		M	F	M	F	M	F	M	F	M	F	M	F
Female	50	0.292	0.708	0.133	0.867	0.249	0.751	0.137	0.863	0.232	0.768	0.201	0.799
	60	0.135	0.865	0.051	0.949	0.093	0.907	0.059	0.941	0.093	0.907	0.090	0.910
	70	0.048	0.952	0.017	0.983	0.027	0.973	0.019	0.981	0.025	0.975	0.034	0.966
	80	0.015	0.985	0.003	0.997	0.008	0.992	0.007	0.993	0.008	0.991	0.010	0.990
	90	0.045	0.955	0.034	0.966	0.030	0.970	0.030	0.970	0.042	0.958	0.036	0.958
Male	50	0.524	0.476	0.303	0.697	0.536	0.464	0.298	0.702	0.529	0.471	0.421	0.579
	60	0.715	0.285	0.516	0.484	0.795	0.205	0.515	0.485	0.789	0.211	0.666	0.334
	70	0.868	0.132	0.747	0.253	0.940	0.060	0.727	0.273	0.942	0.058	0.851	0.149
	80	0.949	0.051	0.903	0.097	0.981	0.019	0.888	0.112	0.981	0.019	0.945	0.055
	90	0.957	0.043	0.931	0.069	0.973	0.028	0.928	0.072	0.974	0.026	0.956	0.044

Table 8.3: We leveraged our bias direction SVC models to predict gender labels of podcast entity vectors. This table shows how these predicted labels change based on the percent of listeners who are male or female in comparison to the type of model used and if gender was (WG) or was not (NG) leveraged during training.

Gender Bias by Genre

Finally, we examine if gender stereotyped genres are more or less likely to be associated with misclassifications of gender if gender is used as a feature or not. This association is evaluated by observing the predicted labels of the sports or true crime podcasts. Results are in Table 8.4.

We found that results for true crime and sports podcasts from SVC and CSVC-2 remained relatively stable when gender was and was not used as a feature during training. When testing for significance, we found that both models did not experience a significant change, with p-values of 0.007 and 0.264, respectively.

However, we found this untrue when testing CSVC-1, which was trained on the 200 "most gender-biased" users. Predictions from CSVC-1 showed a significant change in the precision of predicting true crime podcasts as female, with the metric reducing from 0.80 to 0.49. This drop means more sports podcasts were classified as "female" instead of "male." One can speculate that this reflects AAB for sports podcasts concerning the 200 "most gender-biased" users to have been significantly reduced when removing user gender from the training process. This behavior is also reflected in the significant drop in recall for sports podcasts regarding the CSVC-1 model. In contrast, true crime podcasts experience a slight uptick in the recall, with an increase of 0.81 to 0.84 for CSVC-1 and SVC and CSVC-2 results.

These results show an imbalanced effect of the chosen mitigation method to remove user gender from training. This assumption is further supported when testing for significance between the genre groups of model performance. When gender was used during training, model performance for predicting the stereotyped gender for a podcast was significantly different for all three classification

Bias Direction Model			Podcast Genre	
			Sport	True Crime
			Precision	
	SVC	WG	0.96	0.74
		NG	0.97	0.74
	CSVC-1	WG	0.94	0.80
		NG	0.94	0.49
	CSVC-2	WG	0.95	0.80
		NG	0.95	0.79
			Recall	
	SVC	WG	0.91	0.87
		NG	0.91	0.89
	CSVC-1	WG	0.94	0.81
		NG	0.74	0.84
	CSVC-2	WG	0.94	0.83
		NG	0.93	0.84
			F1-Score	
	SVC	WG	0.93	0.80
		NG	0.94	0.81
	CSVC-1	WG	0.94	0.81
		NG	0.83	0.62
	CSVC-2	WG	0.94	0.82
		NG	0.94	0.81

Table 8.4: Classification performance scores when classifying sport podcasts as "male" or true crime podcasts as "female" when leveraging SVC models used to create bias directions. Acronym descriptions can be found in §6.1.

models. If gender was removed during training, we found that the difference in performance was no longer significant for the SVC model. This difference was not due to the lessening of bias when predicting the gender of a podcast but instead from the classification model predicting more true crime podcasts as "female."

User Gender from Podcast Listening History

We designed this scenario to capture the ability of item embeddings to relay sensitive information about users in downstream models. If item embeddings can be used to predict the user-sensitive attribute, as well as the user embedding itself, it can be assumed that the sensitive attribute is entangled within the item embedding.

We found the overall change in test accuracy to be small when leveraging podcast vectors trained with and without access to gender as a feature. Classification test accuracy for with-gender podcast vectors was 0.832, while non-gender podcast vectors achieved a test accuracy of 0.829. When breaking down results by gender, we found the change in test accuracy to be more pronounced. When gender was included as a feature, 17.9% of female users were classified as male. This percentage reduced to 11.3% when vectors were trained without access to gender. Alternatively, misclassification for male users increased to 23.7% when the model was trained without gender versus 15.7% when trained with user gender as a feature. To better understand the vectors resulting in misclassification, we evaluated the cosine similarity of these vectors against the female, male, female podcast-listening, and male podcast-listening centroids. We found that misclassified podcast vectors showed higher cosine similarity with the opposite gender and gendered listening centroids.

8.2.2 MovieLens Movie Recommendation

We leveraged similar classification scenarios to that designed for Spotify Podcast Recommendation. However, instead of validating against multiple SVC gender directions, we validated against our two modeling scenarios BPR and DMF.

Gendered Movie Watching

In this section, we explore how gender label predictions of movies correspond with gendered listening patterns. Before describing results, it is important to note data modifications we conducted to account for the inherent imbalance in data collected by users according to gender. We calculate the gender percentage of watching for the movies with interactions from randomly sampled 1500 female users and 1500 male users. Additionally, we observe results in relation to the number of interactions for the movie by conducting analysis for all movies, movies with under 50 interactions, and movies with over 50 interactions. This grouping allows us to understand how latent gender bias may relate to the popularity of the movie.

We found that the "popularity" (where we define popularity as the number of times a movie is interacted with in the interaction dataset) and our feedback simulation to significantly change our results in predicting the gender labels for movies. Even in the original DMF embeddings, popularity plays a key position in how the individual movies experience user gender AAB. Table X shows that movies with less than 50 interactions with higher proportions of female listening are not classified "accurately", while movies with more than 50 interactions are much more likely to be increasingly classified as female according to their female listening proportion. This shows that for the DMF algorithm, if a movie is

		DMF					
		1			5		
			# of Interactions			# of Interactions	
		All	$i < 50$	$i > 50$	All	$i < 50$	$i > 50$
		% Predicted Female					
Female Eng. %	50	0.1235	0.0036	0.2075	0.3184	0.0912	0.4759
	60	0.2960	0.0206	0.5588	0.4726	0.1701	0.7500
	70	0.3209	0.1212	0.6349	0.4382	0.1616	0.8730
	80	0.3000	0.0555	0.9231	0.4000	0.1666	1.0000
	90	0.0103	0.0103	-	0.0103	0.0103	-
		% Predicted Male					
Male Eng. %	50	0.9731	0.9928	0.9633	0.8846	0.9391	0.8586
	60	0.9973	0.9962	0.9978	0.9514	0.9584	0.9474
	70	1.0000	1.0000	1.0000	0.9858	0.9946	0.9788
	80	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	90	1.0000	1.0000	-	0.9937	0.9937	-

Table 8.5: This table showcases, for the DMF MovieLens 1M recommendations, how percentage of movies predicted as female versus male change according to their gender engagement percentages and number of interactions. Please note that Eng. is short for Engagement.

more popular and experiences a higher female proportion of users, it is more likely to experience bias towards female users within the recommendation system. These prediction rates are shown to increase after our feedback loop simulation. Our results show that movie embeddings with higher rates of female interactions experience a reinforcing of the bias, with the female prediction rates increasing after five rounds of feedback simulation.

Interestingly, we see the opposite to be true for "male" movies. Even though in both versions of embeddings the male predictions are more "accurate", we can see a decrease in the percentage of movies predicted as male (primarily with male interaction rates between 50 and 70%). This lessening could be attributed to the increase of female "interactions" being captured during training, thus causing an inherent shift in the recommendation training set towards female interactions. This shift may also be a driving factor in the increase in female prediction rates for movies with high female interaction rates, causing both a slight balancing effect for "male" movies and reinforcing effect for "female" movies.

In comparison to DMF results, BPR prediction rates for "female" movies is greater across all movies (regradless of popularity) and simulation version. However, we can see that the results between simulation versions are more stable with prediction rates slightly improving or decreasing for gender proportion groups. We find that the change in rates is correlated with the number of movies within the gender proportion group, with larger groups experiencing a larger increase in reinforcement of user gender bias. The opposite holds true for "male" movies with larger groups experiencing greater dips in prediction rates of male labels. Similar to female movies, male movie BPR embeddings are more stable between feedback loop simulations, experiencing less significant drops than their DMF counterparts. There also seems to be a less drastic difference for results between

		BPR					
		1			5		
			# of Interactions			# of Interactions	
		All	$i < 50$	$i > 50$	All	$i < 50$	$i > 50$
		% Predicted Female					
Female Eng. %	50	0.3943	0.3248	0.4430	0.4434	0.3759	0.4911
	60	0.7487	0.5927	0.9019	0.7661	0.6340	0.8921
	70	0.8765	0.7979	1.0000	0.8827	0.8081	1.0000
	80	0.8000	0.7222	1.0000	0.7800	0.6944	1.0000
	90	0.4639	0.4639	-	0.3917	0.3917	-
		% Predicted Male					
Male Eng. %	50	0.8822	0.7956	0.9249	0.8543	0.7634	0.8970
	60	0.9730	0.9396	0.9915	0.9676	0.9245	0.9915
	70	0.9858	0.9677	1.0000	0.9905	0.9784	1.0000
	80	0.9813	0.9734	1.0000	0.9875	0.9823	1.0000
	90	0.9625	0.9625	-	0.9562	0.9562	-

Table 8.6: This table showcases, for the BPR MovieLens 1M recommendations, how percentage of movies predicted as female versus male change according to their gender engagement percentages and number of interactions. Please note that Eng. is short for Engagement.

	Predicted Female		Predicted Male	
	1	5	1	5
	BPR			
Male Genres	0.0408	0.0356	0.9591	0.9643
Female Genres	0.3561	0.3912	0.6438	0.6087
Neutral Genres	0.2128	0.2307	0.7871	0.7692
	DMF			
Male Genres	0.1280	0.2601	0.8719	0.7398
Female Genres	0.0146	0.0597	0.9853	0.9402
Neutral Genres	0.0847	0.1536	0.9152	0.8463

Table 8.7: This table showcases how different genres of movies are predicted as male or female with respect to the number of simulations and algorithm.

male and female movies for BPR embeddings than DMF embeddings.

Gendered Genres

We found that stereotypically female genres where overwhelmingly more likely to be predicted as female than stereotypically male or neutral genres. In table X, we show case the breakdowns between stereotypically male, female, and neutral genres and how they are predicted between the DMF and BPR models. The original embeddings result in BPR showcasing the strongest signs of user gender bias in the female genre movie embeddings. When simulating the feedback loops, we find that DMF experiences the largest drift with all genres resulting in higher rates of female classification, with the largest change for female genre movies. Unlike DMF, BPR shows signs of distancing between male genres and female genres given the higher rates in stereotypical predictions for both male and female genres after running recommendation simulations.

	DMF		BPR	
	All Users	Sampled	All Users	Sampled
		1		
Predicted as Female	0.0345	0.0684	0.0602	0.0778
		5		
	0.0778	0.4558	0.0614	0.4774

Table 8.8: This table showcases how users are predicted as female based off of their listening history versus the number of simulations, algorithm, and when male users are undersampled when training the bias direction.

Gendered User History

This portion of analysis highlighted the importance of attribute sampling when creating directions, as well as for training in general. We found that female predictions were surprisingly low for user's centroid watching history. Given the analysis in previous sections, we knew that there were significant levels of AAB in accordance with movie genre gender stereotypes, but those findings were not inline with the results we discovered when predicting gender off of user history. Only 66 users were predicted as female based off of their listening history for the DMF SVC model. After further investigation, we found that the our SVC direction and model was overindexing on male users, causing female users to be less accurately represented by the direction we found. If we undersample male users when training the SVC model, we find that the number of female predictions raises significantly.

Table 8.8 shows that this behavior is true for both DMF and BPR algorithms, the extreme increase when accounting for sampling supports the notion that overtime, algorithms are at risk of reinforcing bias within the latent space. This finding raises concerns with our original training methodology leveraging the entire dataset instead of training on equal groups of female and male users.

We investigate how this changes AAB in our mitigation section as a comparison to more advanced mitigation techniques such as adversarial recommendation learning. These results also may point to the discrepancy in performance between female and male stereotyped genres in the previous two sections. Given this was found in the third and final training scenario, revisiting results with under- or over-sampled training data was out of scope for this dissertation. Re-conducting this analysis to observe these differences could be a potential future direction of research.

Chapter 9

Monitor & Flag: Mitigating AAB

In review of results from our evaluation section, we found that both BPR and DMF models showcased significant levels of user gender AAB both via bias metric results and when exploring classification for exploring bias. For this dissertation, we chose to leverage bias metric results to flag where to implement mitigation methods. However, one could leverage classification results in practice as well. We chose to focus on bias metric results because they highlighted a clear front runner for mitigation with the BPR algorithm showcasing the highest levels of AAB when trained on the 1M MovieLens dataset. Given this, and to avoid over-extending the length of this dissertation, we chose to explore mitigation methods only for this recommendation scenario. Future work could explore how mitigation methods perform across different algorithms, but that is out of scope for this dissertation.

9.1 Resampling Training Data

We explored both over- and under- sampling the training data to balance the dataset between male and female users. When undersampling, we found that the performance of the model suffered from loss of information. This was seen from results in our performance metrics, which were significantly less than both the original model and the other mitigated performance metrics. The opposite was true when oversampling the data. By oversampling female user interactions, the model achieved the highest performance scores for MRR, hit, and precision at 10. The other performance metrics, recall and NDCG at 10 were the second highest in comparison to the other model iterations.

Bias direction results remain significant for both the SVC and centroid bias direction. Additionally, we find that bias metric results worsen stereotypically female genre categories for both the under and over sampled training iterations. Stereotypically male genres bias metrics primarily lessen in severity, which is to be expected as these relationships should be reflected less strongly in the training data. Overall, the stereotyped results remain significant with both sampling iterations. However, given the increase in performance metrics oversampling female user data could be seen as beneficial from a performance perspective. This case study could benefit from a multi-step mitigation to optimize for both performance and reducing AAB.

9.2 Iterative Nullspace Projection

From our results, we found that this mitigation method did well at targeting SVC RRIPA results but did not result in the same level of reduction for EAA

results. This finding is relatively intuitive since the mitigation method focuses on removing reliance on the bias direction directly and assumes that bias direction adequately captures the attribute within the latent space. Since EAA results did not experience the same reduction, one can assume that the original bias direction does not fully explain how gender AAB exists between entity vectors. We found that this mitigation method may be best used when the practitioner is highly confident in the bias direction they are mitigating against. EAA metrics having less of an effect from the mitigation may not necessarily be a bad result since it is important to retain the entity relationships within the space. If the historical data leveraged for recommendations is stereotypically biased, it may be difficult to completely eradicate those relationships without potentially exploring two-way relationship mitigation (like Oscar).

However, mitigating for specific two-way relationships (as mentioned previously) may fail to account for non-binary relationships with stereotypical behavior within the latent space. Our goal is to lower AAB across all possible binary relationships between genre and gender, which is satisfied to some extent with iterative nullspace projection. Unfortunately, this mitigation does not target EAA metrics, which would most likely be better targeted with a mitigation method like OsCAR if only one two-way relationship is being targeted. For our mitigation goals, more complicated methods to disentangle gender AAB would be necessary which will be explored in the next section with intrinsic adversarial mitigation methods.

We observe how AAB directions change when mitigated via linear projection on to the SVC or Centroid direction and iterative null projection with 200 or 1000 rounds. The most drastic change when calculating SVC R-RIPA results from projecting onto the centroid bias direction. This scenario results in the lowest

SVC train and test accuracy when predicting gender, with a training accuracy of 0.7204 and training accuracy of 0.7171. It also results in both the centroid and SVC gender bias direction not being statistically significant within the latent space. Average cosine similarity for male users with the trained SVC direction was -0.3428 and for females, -0.3478. Average cosine similarity for male users with the centroid direction was -0.4042 and for females, -0.4106. This renders the bias direction useless for measuring AAB metrics, as showcased by the final results and lack of significant differences between genres.

We look at changes for SVC and Centroid R-RIPA and EAA metrics after mitigating via these four post-processing mitigation methods. We find that out of the four implementations, linear projection on to the Centroid bias direction results in the most noticeable change in the R-RIPA metrics. It is the only method that results in distinct sign changes in the genre bias metrics, with Sci-Fi and War becoming more "female" and all "female" genres becoming more "male". At first, one may think that all genres have become more male, but when you compare against the negative cosine similarities of the users with these unsignificant bias directions, it may be more conducive of a fundamental change in the latent relationships themselves with user and item entity vectors becoming less entangled within the space. Additionally, since the bias directions are no longer significant, the bias metrics should be seen as not significant as well. One can see that this does not occur when projecting onto the SVC direction. This method results in less drastic changes across the various genres and maintains some level of the stereotypical associations between genres and gender. Iterative Null Projection is found to be much more fruitful than basic linear projection. However, we can see that mitigation gains taper with an increase in iterative rounds.

If we compare against the performance of the new mitigated vectors, in addition to achieving non-significant bias directions and metrics, projecting onto the centroid bias direction also results in high levels of performance. The best performance between these methods is projecting once onto the SVC bias direction, but as we can see in tables x and y, it does not achieve the same reduction in AAB and retains significant bias directions. Iterative Null Projection results in similar levels of performance to one another, but a noticeable drop from one-time projection mitigation. We find that the linear projection mitigation method also raises performance metrics from the original baseline, showcasing that reinforcing stereotypes via AAB in the latent space may decrease user experience. But given the fact that SVC linear projection achieves the highest performance but not the largest decrease in AAB signals that retaining some level of AAB could be beneficial, thus allowing to mimic some levels of stereotypical user behaviors while allowing for more diversity in the recommendations which has been proven to increase the success of recommendations.

9.3 Adversarial Recommendation for BPR

In comparison to iterative nullspace projection, we found adversarial recommendation with attribute protection (RAP) to better mitigate a wider range of AAB metrics. However, we found that performance metrics were significantly diminished from original levels and in comparison to simple one step null projection onto the centroid or SVC bias direction. This trade-off is a known problem within the bias mitigation research space, and determining the correct level of trade-off is beyond the scope of this dissertation.

A particularly interesting result was found in our bias direction results. The

user entity vectors from the adversarial RAP resulted in complete linear separation between male and female users (as shown in table X). This resulted in a more statistically significant bias direction according to gender. Contrastingly, with the more significant bias direction, most bias attribute association metrics significantly decreased, meaning specific item groups were less stereotypically associated with specific user genders. Even though bias levels dropped, stereotypical differences still remained. Given the recommendations reflect real user behavior, eradicating stereotypes from predictions may not be realistic and could even be seen as detrimental for users who do have more stereotypical engagement patterns. The results from adversarial mitigation shows that one can reduce AAB to less harmful levels while maintaining performance, thus making this method a potential starting point for mitigating this category of recommendation system bias. Future work could include exploring why the adversarial component strengthens the linear separation between users by gender and how iterations of inputs into the adversarial component affect AAB metrics. Additionally, one could experiment with implementing multiple components of mitigation in attempt to address multiple bias metrics while maintaining higher levels of performance.

When comparing with post-processing techniques, adversarial mitigation primarily achieved the best mitigation results for stereotypically female genres. Unlike null projection methods, adversarial mitigation was able to achieve favorable results for both SVC and Centroid R-RIPA, showcasing that this method is more flexible to addressing multiple AAB metrics. However, as shown in table X, this does not hold true for EAA results, where we see EAA increasing for stereotypically male genres. Unlike R-RIPA, EAA is reflective of all possible relationships between each attribute and test entity. If we take into account the fact that

		Recall	MRR	NDCG	Hit	Precision
Under		0.0914	0.1413	0.0838	0.3307	0.0529
Over		0.1604	0.4794	0.2878	0.8098	0.2441
Linear	SVC	0.1900	0.4910	0.2925	0.7937	0.2270
Projection	Centroid	0.1788	0.4757	0.2760	0.7722	0.2123
Iterative Null	n=200	0.1282	0.3563	0.1891	0.6806	0.1473
Projection	n=1000	0.1242	0.3490	0.1901	0.6632	0.1516
Adversarial BPR		0.1462	0.3667	0.2000	0.7248	0.1573
BPR		0.1464	0.3770	0.2052	0.7267	0.1608

Table 9.1: Performance metrics against the mitigation method leveraged. We can see that the performance metrics increase when we mitigate for user gender bias by oversampling female users during training and for both simple linear projection mitigation implementations. The abbreviation Under reflects under sampling male users. The abbreviation Over reflects over sampling female users.

the data is skewed towards male users, one could hypothesize that this result is indicative of adversarial mitigation forcing all movie entities to be more "male". This behavior would be beneficial to the adversarial algorithm because it would inherently make it more difficult to predict the gender of a user and item pairing. Table X shows that performance metrics are relatively unchanged between the adversarial and original BPR model, which could indicate that even though there are changes in AAB metrics, stereotypes in the final results may be unchanged.

	BPR	Under-sample Male Users	Over-sample Female Users	Adversarial BPR
			SVC R-RIPA	
Action	0.1677	0.1682	0.1446	0.1159
Sci-Fi	0.1601	0.1363	0.1583	0.0969
War	0.2421	0.1626	0.1532	0.1902
Western	0.2164	0.1482	0.1117	0.1513
Crime	0.0928	0.0795	0.0494	0.0641
Romance	-0.1879	-0.2016	-0.2059	-0.0764
Children's	-0.0867	-0.1208	-0.1204	-0.0162
Drama	-0.0712	-0.1227	-0.1680	-0.0157
			Centroid R-RIPA	
Action	0.3300	0.3750	0.3000	0.2161
Sci-Fi	0.3222	0.3668	0.3076	0.2097
War	0.2732	0.2716	0.2348	0.1902
Western	0.2107	0.2224	0.1785	0.1513
Crime	0.1763	0.1749	0.1375	0.1265
Romance	-0.3225	-0.3404	-0.2887	-0.1520
Children's	-0.1168	-0.1356	-0.1135	-0.0397
Drama	-0.1703	-0.2213	-0.1969	-0.0555
			GEAA	
Action	31.78	35.56	28.56	39.11
Sci-Fi	20.34	23.41	19.19	24.88
War	4.05	4.09	3.44	6.93
Western	2.90	2.99	2.43	5.31
Crime	5.27	5.22	4.07	7.12
Romance	-34.37	-36.26	-30.41	-30.44
Children's	-6.99	-8.24	-6.72	-4.47
Drama	-56.87	-72.48	-65.02	-34.87

Table 9.2: Bias metrics reflected against the mitigation methods: over sampling female users, under sampling male users, and Adversarial BPR.

	BPR	Linear Projection		Iterative Null Projection	
		SVC	Centroid	n=200	n=1000
			SVC R-RIPA		
Action	0.1677	0.1400	0.0069	0.0596	0.0499
Sci-Fi	0.1601	0.1439	-0.0214	0.0565	0.0462
War	0.2421	0.1160	-0.0079	0.0907	0.0718
Western	0.2164	0.1040	0.0042	0.0742	0.0676
Crime	0.0928	0.0994	0.0694	0.0298	0.0279
Romance	-0.1879	-0.1834	0.0210	-0.0680	-0.0556
Children's	-0.0867	-0.0707	0.0782	-0.0297	-0.0249
Drama	-0.0712	-0.0767	0.1557	-0.0273	-0.0210
			Centroid R-RIPA		
Action	0.3300	0.2991	0.0329	0.3302	0.3420
Sci-Fi	0.3222	0.2957	0.0960	0.2743	0.4106
War	0.2732	0.1160	-0.0151	0.2822	0.3259
Western	0.2107	0.0945	0.0042	0.2617	0.1755
Crime	0.1763	0.0994	0.0411	0.0261	0.1897
Romance	-0.3225	-0.1834	0.0444	-0.2857	-0.3041
Children's	-0.1168	-0.0795	0.1262	-0.0534	-0.2199
Drama	-0.1703	-0.0767	0.1557	-0.0227	-0.1409
			GEAA		
Action	31.78	21.87	0.07	28.30	37.30
Sci-Fi	20.34	14.17	0.15	15.46	39.36
War	4.05	1.79	-0.01	3.73	5.47
Western	2.90	0.98	0.01	3.21	2.73
Crime	5.27	3.59	0.03	3.96	6.43
Romance	-34.37	22.09	0.11	-27.17	-36.71
Children's	-6.99	-3.62	0.18	-2.85	-14.92
Drama	-56.87	-41.69	1.08	-38.69	-53.29

Table 9.3: Bias metrics reflected against the mitigation methods: SVC Linear Projection, Centroid Linear Projection, 200 rounds of Iterative Null Projection, and 1000 rounds of Iterative Null Projection.

Part III

Conclusion

Chapter 10

Conclusion

In this dissertation, we introduce a variety of frameworks and methods to aid practioners and academics in the auditing and evaluation of attribute association bias (AAB) in their recommendation systems. We lay the scene by proposing a wider framework for disaggregated audits of recommendation systems in practice. This framework, SIIM, provides a framework for our analysis methods of AAB. Following the introduction of SIIM, we propose the definition of AAB and present the novel ESA framework to evaluate AAB in practice. The ESA framework consists of three steps which look to understand the existence, significance, and amplification of attribute association bias. For each step of the ESA framework, we provide various methods to complete the corresponding step in practice. In the first step, we explore how to understand the existence of the attribute in the latent space as defined by the attribute defining entity steps. We do this by calculating attribute bias directions. The second step looks to evaluate the significance of relationships between non-attribute defining entities in the test sets and the bias directions, as well as the attribute association bias direction in the space. Finally, we explore the amplification of the attribute association

bias by evaluating feedback loops and classification scenarios concerning how the bias may change over time and manifest within non-attribute related entity vectors. After implementing the ESA framework, we explore mitigation methods for AAB leveraging adversarial recommendation techniques. In order to implement these proposed methods and frameworks, we leverage two datasets, a proprietary dataset from Spotify for podcast recommendations and MovieLens.

Our framework provides a clear path in uncovering potentially harmful stereotyped relationships resulting from AAB resulting from an LFR model. In showcasing our techniques, we found that our proposed methodologies successfully measured and flagged AAB. Additionally, we uncovered clear advantages and disadvantages for our proposed methods to help practitioners choose the appropriate techniques for their scoped evaluations. In the following sections, we would like to address and discuss case study specific results, limitations to our research, and potential future directions.

10.1 Gender in Latent Factor Recommendation

Throughout this dissertation, user gender was the main sensitive attribute driving our research in AAB in latent factor recommendation models. Thus, it is only fitting that we discuss our findings on the effects of user gender on recommendations. Our findings support the idea that systematic bias occurs as AAB in LFR outputs, which leads us to a more challenging question: when is it appropriate to mitigate for systematic bias? In some cases, like in our case studies, stereotyped behavior is common and could even be seen as beneficial for providing useful recommendations to users. If the existence of implicit AAB actually improves user experience, how should one reduce the risk of representative harms?

In the case of user gender bias, the harm lies in the model potentially reinforcing stereotypes by driving users towards gendered listening habits. It is possible to monitor levels of AAB overtime in order to flag increasing bias in the latent space. But when are levels of reinforcement considered harmful? Both the research and practitioner community would benefit from more exploration of how to approach setting baselines for managing representative harms in recommendations.

10.1.1 Mitigating Gender AAB

[41] suggested the capability for user gender bias to be systematic bias embedded within the latent space, thus making it difficult for simple mitigation techniques to address the core issue at hand. Their study demonstrated "a systematic bias found in the embeddings, which is independent of the gender direction" [41]. Given this independence, debiasing methods grounded in removing the gender direction were found to be "superficial" fixes. Systematic bias in our case study, similar in nature to that found in word embeddings, would result in AAB remaining significant even when user gender is not leveraged as a model feature. In mitigating user gender AAB with similar methods of null space projection, we found the same to be true for LFR algorithms. Removing gender completely from the trained latent space is next to impossible, and this removal could actually incur harmful performance results for users as well.

In context-aware podcast recommendation, we found that removing user gender as a feature resulted in a statistically significant decrease in levels of AAB, however, significant implicit AAB remained. This finding was true for our movie recommendation case study for all three mitigation methods. Like [41] 's results,

our case study observations suggest that gender stereotypes can become implicitly embedded in the representations of both users and items, as supported by the persistence of this bias when gender is not explicitly used as a feature and when non-context aware recommendation models are mitigated. The presence of implicit AAB signals the potential for systematic gender bias when leveraging latent factor recommendation models, or potentially recommendation algorithms in general, when users historically show stereotyped patterns. This finding was not surprising given previous research detailing the highly gendered nature of podcast listening and movie watching[19, 27, 80]. Our findings leveraging our framework demonstrate that, similar to [41], known systematic bias can be found and quantified in recommendations. Given this, it is essential for practitioners to audit for AAB when systematic bias is a known factor in their recommendation scenario, such as podcast or movie recommendations.

Additionally, we find that mitigating gender AAB in our BPR model for movie recommendation did not completely remove the bias. When a method did reduce bias more, we found that performance experienced more adverse effects. This result further supports the fact that mitigating bias and maintaining performance is a trade-off that needs to be addressed in practice. Defining trade-offs and thresholds is a subjective and difficult process which is most often project specific. There is little guidance in how best to approach this common problem and could be an impactful area for future research.

Even though standard performance metrics were negatively impacted, other evaluation metrics such as those addressing diversity and novelty may improve when reducing bias. Future work could include exploring how reducing stereotyping bias, such as AAB, leads to more diverse and positive experiences for recommendation stakeholders.

10.2 Binary versus Multi-categorical Metrics

In this dissertation, we proposed evaluation approaches for both binary and multi-categorical group situations. While exploring and creating these methods, we found that the intense need, expressed in both academic and industry settings, for a holistic multi-categorical metric may not result in as much impact as expected. It is extremely difficult to capture the nuances of multi-categorical relationships and behaviors in one holistic metric due to the inherent masking of results that comes with aggregated metrics (simple examples being, mean and median). This challenge becomes even more noticeable as the number of groups increase, as a result, even with a holistic metric, pairwise and binary comparisons may be necessary to make informed decisions regarding bias evaluation and mitigation.

Our multi-categorical bias direction method proved helpful for leveled categorical groups (such as temperature and heat). However, leveraging our simple method may be unfruitful for capturing holistic behaviors of non-scale related categories (such as occupation and race). One could look to explore the creation of a holistic bias direction for those types of groups, but the methods would be significantly more complicated, difficult to understand, and most likely challenging to implement in practice. This creates the question of do we actually need one all-encompassing metric, or is it sufficient to settle for pairwise techniques when group-specific exploration will be required anyways? Additionally, is it truly good practice to leverage information masking aggregate metrics when they can lead to missing group-specific harmful bias?

A holistic aggregate bias metric may be helpful in theory, but one needs to be aware that this area of research has the potential to affect people's livelihoods. Is it worth it to risk missing harm in favor of reducing the computational overhead

of pairwise comparisons? This trade-off could benefit from closer investigation to help inform researchers and practitioners when a holistic aggregate bias metric is truly beneficial to their workflow.

10.3 Limitations & Future Work

Our most noticeable limitation when implementing these techniques was the lack of distinct and well-labeled user and item pairings for metric calculation, a common occurrence in recommendation settings. We demonstrated methods to overcome this limitation, but this work could be further refined in the future to avoid possibly introducing more bias into the evaluation via practitioner-defined entity pairing techniques. In the future, specific to the podcast recommendation case study, we plan to explore counterfactual user vectors according to gender to create distinct pairs. Counterfactual user pairings would isolate the feature within the latent space and potentially reduce attributing spurious relationships between users solely to gender differences. It is important to note that this workaround is only available for models trained with entity attributes. This limitation would remain when evaluating the implicit or systematic bias of an LFR algorithm.

10.3.1 Practical Limitations

Despite the proliferation of work focusing on algorithmic auditing, we've encountered numerous challenges in evaluating bias in practice. A lack of standards and guidance leaves practitioners with significant challenges, which go beyond tooling not always being suitable [74] and beyond conceptual tensions in different definitions of fairness [44]. The challenges we encountered in practice are likely shared across many organizations, particularly as internal audits become

more common. Practitioners may be responsible for developing playbooks of instructions to audit and monitor systems, and would benefit from sharing challenges and lessons learned. We hope that by sharing our own challenges with the community, we can uncover shared obstacles and work collaboratively between industry and academia to ensure best practices for tackling the complex task of evaluating recommendation systems.

Addressing Bias Thresholds It is important to note that determining these thresholds is inherently challenging due to the gap between "practical" and "statistical" for assessing algorithmic impact [13]. Statistical significance is influenced by many decisions (e.g., alpha, one- vs. two-sided test) that are generally designed to assess whether results occurred due to chance, not whether the results were meaningful [13]. "Significant" does not mean practical, and "marginally better than before" might not be very impactful [13]. In the case of mitigation research, determining these types of thresholds would require researchers to make more definitive statements of what is or is not fair. Making these decisions is difficult due to the lack of standards in the space and the need for subject matter expertise to understand the nuances of fairness within the studied domain. However, by not researching or making these difficult decisions, this responsibility falls upon industry practitioners, leaving them at risk of making sub-optimal decisions leading to potentially harmful downstream effects on providers and consumers. There is an incredible opportunity to increase the impact of this type of mitigation work by researching best practices in approaching the complex subject of thresholds for fairness.

10.3.2 Embedding Functions

As shown in section X observing Bias Directions, understanding how the entities are embedded into the space is paramount to leveraging this evaluation framework correctly. Out of the three algorithms we observed, NCF was the one framework that did not leverage an embedding function allowing for the model to embed the user and item entities into the same space. We demonstrated how that led to confusing results when testing for significance against the other two algorithms (which is to be expected). If the algorithm results in embeddings, it is not guaranteed that it embeds users and items into the same space. Our evaluation framework only works for latent factor recommendation algorithms that result in one user-item embedding space, such as BPR and DMF.

10.4 Concluding Remarks

The success of our methodologies in uncovering AAB highlights the importance of understanding how stereotypical relationships can become embedded into trained recommendation latent spaces. For example, our ability to predict user gender from podcast vectors demonstrates how leveraging these vectors as attributes in downstream models can introduce implicit user gender bias in subsequent outputs, even if owners of downstream models intentionally remove user gender as a training feature. The ability for listening history to predict user gender showcases that user gender bias is embedded within the podcast vectors, meaning their use can inherently introduce gender bias into other modeling systems. Understanding this type of representation bias becomes increasingly crucial in industry recommendation systems where embeddings are used across models owned by different teams.

For example, if a team audits and mitigates its model for user gender bias but leverages said podcast vectors as a feature, any unrelated models leveraging said feature could be unknowingly reintroduced to user gender bias. If AAB is left unchecked in hybrid recommendation scenarios, teams are at risk of amplifying systematic representation harms resulting from providing stereotyped recommendations for stakeholders. Similar to findings by [12], our results support the notion that capturing and understanding the behavior of gender bias in more implicitly biased recommendation vector embeddings is a complicated and nuanced task, requiring further analysis beyond our results showcased in this paper. We hope our evaluation framework serves as a building block for future research addressing representative harms and AAB in recommendation systems.

Bibliography

[1] The right pitch: A look into the popularity of podcast hosts by gender.

[2] Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J., and Pizzato, L. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction 30*, 1 (2020), 127–158.

[3] Abdollahpouri, H., and Burke, R. Reducing Popularity Bias in Recommendation Over Time, June 2019. Number: arXiv:1906.11711 arXiv:1906.11711 [cs].

[4] Abdollahpouri, H., Burke, R., and Mobasher, B. Value-Aware Item Weighting for Long-Tail Recommendation.

[5] Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B., and Malthouse, E. User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (2021), pp. 119–129.

[6] Adomavicius, G., and Tuzhilin, A. Context-aware recommender systems. In *Recommender systems handbook*. Springer, 2011, pp. 217–253.

[7] Alexander, L. What makes wrongful discrimination wrong? biases, preferences, stereotypes, and proxies. *University of Pennsylvania Law Review 141*, 1 (1992), 149–219.

[8] Amatriain, X., and Basilico, J. Recommender systems in industry: A netflix case study. In *Recommender systems handbook*. Springer, 2015, pp. 385–419.

[9] Bakalar, C., Barreto, R., Bergman, S., Bogen, M., Chern, B., Corbett-Davies, S., Hall, M., Kloumann, I., Lam, M., Candela, J. Q., et al. Fairness on the ground: Applying algorithmic fairness approaches to production systems. *arXiv preprint abs/2103.06172* (2021).

[10] Barocas, S., Crawford, K., Shapiro, A., and Wallach, H. The problem with bias: Allocative versus representational harms in machine learning, 2017.

[11] Barocas, S., Guo, A., Kamar, E., Krones, J., Morris, M. R., Vaughan, J. W., Wadsworth, W. D., and Wallach, H. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (2021), pp. 368–378.

[12] Basta, C., Costa-Jussà, M. R., and Casas, N. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783* (2019).

[13] Beattie, L., Taber, D., and Cramer, H. Challenges in translating research to practice for evaluating fairness and bias in recommendation systems. In *Proceedings of the 16th ACM Conference on Recommender Systems* (2022), pp. 528–530.

[14] Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., and Goodrow, C. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (New York, NY, USA, 2019), KDD '19, Association for Computing Machinery, p. 2212–2220.

[15] Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. Data decisions and theoretical implications when adversarially learning fair representations.

[16] Bhardwaj, R., Majumder, N., and Poria, S. Investigating gender bias in bert. *Cognitive Computation 13*, 4 (2021), 1008–1018.

[17] Biega, A. J., Gummadi, K. P., and Weikum, G. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (New York, NY, USA, 2018), SIGIR '18, Association for Computing Machinery, p. 405–414.

[18] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. Fairlearn: A toolkit for assessing and improving fairness in ai. Tech. Rep. MSR-TR-2020-32, Microsoft, May 2020.

[19] Boling, K. S., and Hull, K. Undisclosed information—serial is my favorite murder: Examining motivations in the true crime podcast audience. *Journal of Radio & Audio Media 25*, 1 (2018), 92–108.

[20] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems 29* (2016).

[21] Burke, R. Hybrid web recommender systems. *The adaptive web* (2007), 377–408.

[22] Burke, R. Multisided fairness for recommendation, 2017.

[23] Burke, R. D., Abdollahpouri, H., Mobasher, B., and Gupta, T. Towards multi-stakeholder utility evaluation of recommender systems. *UMAP (Extended Proceedings) 750* (2016).

[24] Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science 356*, 6334 (2017), 183–186.

[25] Chen, L., Ma, R., Hannák, A., and Wilson, C. *Investigating the Impact of Gender on Rank in Resume Search Engines.* Association for Computing Machinery, New York, NY, USA, 2018, p. 1–14.

[26] Covington, P., Adams, J., and Sargin, E. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (New York, NY, USA, 2016), RecSys '16, Association for Computing Machinery, p. 191–198.

[27] Craig, C. M., Brooks, M. E., and Bichard, S. Podcasting on purpose: Exploring motivations for podcast use among young adults. *International Journal of Listening* (2021), 1–10.

[28] Dash, A., Chakraborty, A., Ghosh, S., Mukherjee, A., and Gummadi, K. P. When the umpire is also a player: Bias in private label product recommendations on e-commerce marketplaces. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2021), FAccT '21, Association for Computing Machinery, p. 873–884.

[29] De Divitiis, L., Becattini, F., Baecchi, C., and Bimbo, A. D. Disentangling features for fashion recommendation. *ACM Trans. Multimedia Comput. Commun. Appl.* (apr 2022). Just Accepted.

[30] Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation 10*, 7 (1998), 1895–1923.

[31] Du, Y., Fang, Q., and Nguyen, D. Assessing the reliability of word embedding gender bias measures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021), pp. 10012–10034.

[32] Ekstrand, M. D., Das, A., Burke, R., and Diaz, F. Fairness and discrimination in information access systems, 2021.

[33] Ekstrand, M. D., Das, A., Burke, R., Diaz, F., et al. Fairness in information access systems. *Foundations and Trends® in Information Retrieval 16*, 1-2 (2022), 1–177.

[34] Ekstrand, M. D., Tian, M., Kazi, M. R. I., Mehrpouyan, H., and Kluver, D. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems* (New York, NY, USA, 2018), RecSys '18, Association for Computing Machinery, p. 242–250.

[35] Epps-Darling, A., Bouyer, R. T., and Cramer, H. Artist gender representation in music streaming. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR 2020). ISMIR* (Montréal, Canada, 2020), pp. 248–254.

[36] Ethayarajh, K., Duvenaud, D., and Hirst, G. Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361* (2019).

[37] European Parliament. Directorate General for Parliamentary Research Services. *A governance framework for algorithmic accountability and transparency.* Publications Office, LU, 2019.

[38] Feldman, T., and Peake, A. End-to-end bias mitigation: Removing gender bias in deep learning. *arXiv preprint arXiv:2104.02532* (2021).

[39] Geyik, S. C., Ambler, S., and Kenthapadi, K. Fairness-aware ranking in search &; recommendation systems with application to linkedin talent search. KDD '19, Association for Computing Machinery, p. 2221–2231.

[40] Geyik, S. C., Ambler, S., and Kenthapadi, K. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (New York, NY, USA, 2019), KDD '19, Association for Computing Machinery, p. 2221–2231.

[41] Gonen, H., and Goldberg, Y. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862* (2019).

[42] Götting, M. C. Podcasts in the uk - statistics &; facts, Feb 2022.

[43] Infortuna, C., Battaglia, F., Freedberg, D., Mento, C., Zoccali, R. A., Muscatello, M. R. A., and Bruno, A. The inner muses: How affective temperament traits, gender and age predict film genre preference. *Personality and Individual Differences 178* (2021), 110877.

[44] Jacobs, A. Z., and Wallach, H. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2021), FAccT '21, Association for Computing Machinery, p. 375–385.

[45] Jobin, A., Ienca, M., and Vayena, E. The global landscape of ai ethics guidelines. *Nature Machine Intelligence 1*, 9 (2019), 389–399.

[46] Koren, Y., Rendle, S., and Bell, R. Advances in collaborative filtering. *Recommender systems handbook* (2022), 91–142.

[47] Krcmar, M., and Kean, L. G. Uses and gratifications of media violence: Personality correlates of viewing and liking violent genres. *Media Psychology 7*, 4 (2005), 399–420.

[48] Kuhlman, C., VanValkenburg, M., and Rundensteiner, E. Fare: Diagnostics for fair ranking using pairwise error metrics. In *The World Wide Web Conference* (New York, NY, USA, 2019), WWW '19, Association for Computing Machinery, p. 2936–2942.

[49] Lauscher, A., Glavaš, G., Ponzetto, S. P., and Vulić, I. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), vol. 34, pp. 8131–8138.

[50] Lazovick, M. Women podcast listeners: closing the listening gender gap, 2022.

[51] Lee, M. S. A., and Singh, J. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2021), CHI '21, Association for Computing Machinery.

[52] Liang, P. P., Li, I. M., Zheng, E., Lim, Y. C., Salakhutdinov, R., and Morency, L.-P. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100* (2020).

[53] Liu, W., Guo, J., Sonboli, N., Burke, R., and Zhang, S. Personalized fairness-aware re-ranking for microlending. In *Proceedings of the 13th ACM Conference on Recommender Systems* (2019), pp. 467–471.

[54] Liu, Y., Jun, E., Li, Q., and Heer, J. Latent space cartography: Visual analysis of vector space embeddings. In *Computer Graphics Forum* (2019), vol. 38, Wiley Online Library, pp. 67–78.

[55] Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., and Burke, R. Feedback loop and bias amplification in recommender systems. 2145–2148.

[56] Manzini, T., Lim, Y. C., Tsvetkov, Y., and Black, A. W. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047* (2019).

[57] May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (2019).

[58] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR) 54*, 6 (2021), 1–35.

[59] Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., and Diaz, F. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (New York, NY, USA, 2018), CIKM '18, Association for Computing Machinery, p. 2243–2251.

[60] Melchiorre, A. B., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O., and Schedl, M. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management 58*, 5 (2021), 102666.

[61] Musto, C., Gemmis, M. d., Lops, P., Narducci, F., and Semeraro, G. Semantics and content-based recommendations. In *Recommender systems handbook*. Springer, 2022, pp. 251–298.

[62] Narayanan, A. Translation tutorial: 21 fairness definitions and their politics, 2018.

[63] Nazari, Z., Charbuillet, C., Pages, J., Laurent, M., Charrier, D., Vecchione, B., and Carterette, B. Recommending podcasts for cold-start users based on music listening and taste. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2020), SIGIR '20, Association for Computing Machinery, p. 1041–1050.

[64] Nema, P., Karatzoglou, A., and Radlinski, F. Disentangling preference representations for recommendation critiquing with ß-vae. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (2021), pp. 1356–1365.

[65] of Labor, U. D. Practical Significance in EEO Analysis Frequently Asked Questions | U.S. Department of Labor, 2021.

[66] Oliver, M. B. The respondent gender gap. *Media entertainment: The psychology of its appeal* (2000), 215–234.

[67] Omrani Sabbaghi, S., and Caliskan, A. Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (2022), pp. 518–531.

[68] Orgad, H., Goldfarb-Tarrant, S., and Belinkov, Y. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2022).

[69] Raj, A., and Ekstrand, M. D. Fire dragon and unicorn princess; gender stereotypes and children's products in search engine responses. *arXiv preprint arXiv:2206.13747* (2022).

[70] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (2020), pp. 33–44.

[71] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2020), FAT* '20, Association for Computing Machinery, p. 33–44.

[72] Ramakrishna, A., Malandrakis, N., Staruk, E., and Narayanan, S. A quantitative analysis of gender differences in movies using psycholinguistic normatives. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (2015), pp. 1996–2001.

[73] Ricci, F., Rokach, L., and Shapira, B. Recommender systems: Techniques, applications, and challenges. *Recommender Systems Handbook* (2022), 1–35.

[74] Richardson, B., Garcia-Gathright, J., Way, S. F., Thom, J., and Cramer, H. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ml toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2021), CHI '21, Association for Computing Machinery.

[75] Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., and Turchi, M. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics 9* (08 2021), 845–874.

[76] Shani, G., and Gunawardana, A. Evaluating recommendation systems. In *Recommender systems handbook*. Springer, 2011, pp. 257–297.

[77] Singh, A., and Joachims, T. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (New York, NY, USA, 2018), KDD '18, Association for Computing Machinery, p. 2219–2228.

[78] Sonboli, N., Burke, R., Liu, Z., and Mansoury, M. Fairness-aware recommendation with librec-auto. In *Fourteenth ACM Conference on Recommender Systems* (2020), pp. 594–596.

[79] Sonboli, N., Eskandanian, F., Burke, R., Liu, W., and Mobasher, B. Opportunistic multi-aspect fairness through personalized re-ranking. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (2020), pp. 239–247.

[80] Soto-Vásquez, A. D., Vilceanu, M. O., and Johnson, K. C. "just hanging with my friends": Us latina/o/x perspectives on parasocial relationships in podcast listening during covid-19. *Popular Communication 20*, 4 (2022), 324–337.

[81] Steck, H. Item popularity and recommendation accuracy. In *Proceedings of the fifth ACM conference on Recommender systems* (New York, NY, USA, Oct. 2011), RecSys '11, Association for Computing Machinery, pp. 125–132.

[82] Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).

[83] Tan, Y. C., and Celis, L. E. Assessing social and intersectional biases in contextualized word representations. *Advances in Neural Information Processing Systems 32* (2019).

[84] The Artificial Intelligence Channel. The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford #NIPS2017, December 2017.

[85] Tsintzou, V., Pitoura, E., and Tsaparas, P. Bias disparity in recommendation systems. *arXiv preprint arXiv:1811.01461* (2018).

[86] Wang, H., and Zhang, H. Movie genre preference prediction using machine learning for customer-based information. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)* (2018), IEEE, pp. 110–116.

[87] Wang, X., Li, Q., Yu, D., Cui, P., Wang, Z., and Xu, G. Causal disentanglement for semantics-aware intent learning in recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[88] Watkins, E. A., McKenna, M., and Chen, J. The four-fifths rule is not disparate impact: A woeful tale of epistemic trespassing in algorithmic fairness. *arXiv preprint arXiv:2202.09519* (2022).

[89] Wühr, P., Lange, B. P., and Schwarz, S. Tears or fears? comparing gender stereotypes about movie preferences to actual preferences. *Frontiers in psychology 8* (2017), 428.

[90] Wühr, P., Lange, B. P., and Schwarz, S. Tears or fears? comparing gender stereotypes about movie preferences to actual preferences. *Frontiers in Psychology 8* (2017).

[91] Yang, K., and Stoyanovich, J. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (New York, NY, USA, 2017), SSDBM '17, Association for Computing Machinery.

[92] Yao, S., and Huang, B. Beyond Parity: Fairness Objectives for Collaborative Filtering, Nov. 2017. arXiv: 1705.08804.

[93] Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (2017), pp. 1569–1578.

[94] Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (New York, NY, USA, 2017), CIKM '17, Association for Computing Machinery, p. 1569–1578.

[95] Zhang, H., Sneyd, A., and Stevenson, M. Robustness and reliability of gender bias assessment in word embeddings: The role of base pairs. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (2020), pp. 759–769.

[96] Zhang, S., Tay, Y., Yao, L., Sun, A., and Zhang, C. Deep learning for recommender systems. In *Recommender Systems Handbook*. Springer, 2022, pp. 173–210.

[97] Zhang, Y., Zhu, Z., He, Y., and Caverlee, J. Content-collaborative disentanglement representation learning for enhanced recommendation. In *Fourteenth ACM Conference on Recommender Systems* (2020), pp. 43–52.

[98] Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., and Chang, K.-W. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310* (2019).

[99] Zhao, Z., Chen, J., Zhou, S., He, X., Cao, X., Zhang, F., and Wu, W. Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[100] Zheng, Y., Gao, C., Li, X., He, X., Li, Y., and Jin, D. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021* (2021), pp. 2980–2991.

[101] Zhou, P., Shi, W., Zhao, J., Huang, K.-H., Chen, M., Cotterell, R., and Chang, K.-W. Examining gender bias in languages with grammatical gender. *arXiv preprint arXiv:1909.02224* (2019).