Bachelor thesis

Artificial Intelligence

What do you want to want?

Simulating User-Directed Preference Change in Recommender Systems

Author:
Clara Maine
s1032005

First supervisor:
Prof. Martha Larson
Institute for Computing
and Information Sciences
martha.larson@ru.nl

Daily supervisor:
Manel Skolom
Institute for Computing
and Information Sciences
manel.skolom@ru.nl

Second supervisor:
Dr. T. Kachman
Artificial Intelligence
Radboud University
tal.kachman@donders.ru.nl

June 25, 2022

Abstract

In this thesis we reviewed the problem of recommender system-induced preference change and presented user-directed preference change as a solution. We argue that a reliance on behavior to infer user preferences can lead to gaps and problems in the kind of preference change they tend to invoke in users. User-directed preference change is a bottom-up approach which better supports users to be in-the-loop of their own preference change. In order to test the system-wide effects of user-directed preference change, we performed a simulation study to model the effect of a novel "meta-preference" mechanism by which users can provide corrective feedback to better align recommended content with their preference change goals. Using the T-RECS simulation tool, we formalized user meta-preferences and present a simple mechanism by which they were incorporated into the predicted user preferences of simulated content-based recommender systems. In measuring the degree and alignment of these changed preferences, we confirmed that our mechanism was effective in causing user-directed preference change within the simulation. We noted a trade-off with accuracy and found strong homogenizing effects on recommendations and preferences throughout the simulation. Though we believe that user-directed preference change is an important consideration, we do not consider it to be a complete solution to the preference change problem given its myopic focus on individual satisfaction. Rather, this thesis explores one of many possible mechanisms which are important in the development of technology which can be symbiotically beneficial for individuals and society.

Contents

Chapter 1

Introduction

Moving throughout the world, we encounter recommender systems pervasively. Browsing through social media, we are recommended posts and accounts we may not have chosen to follow, and on streaming platforms we are given curated lists of film and media recommendations. When searching for an item on an online shop, we are shown recommendations in form of added lists of “related products” or “customers also bought”. These are items we did not search for and may not have even known we wanted, but through recommendation our preferences are learned and even eventually changed.

Generally, recommender systems make predictions about future behavior by finding and exploiting patterns in past activity. In doing so, they also influence future behavior by making some choices more likely than others. While behavior is thought to be dependent on some underlying preferences, preferences are crucially not independent from behavior [Ashton and Franklin, 2022]. The two are casually linked, which means that recommender systems do more than simply predicting our preferences, they also play a vital role in shaping them.

This leads us to an important question: if recommender systems are changing the preferences of their users, *how* and *in what ways* are those preferences being changed? Despite the lack of comprehensive research, we are able to observe induced preference change in cases where the effects prove especially disruptive. Social media algorithms have been shown to radicalize and polarize users in the interest of optimizing engagement [Rathje et al., 2021], as well as promote addictive behavior in users [Hasan et al., 2018], sometimes as the an intentional goal [Seaver, 2019].

Beyond these extreme examples, the effect of recommender systems on user preferences is concerningly not known despite the fact that recommender systems have been deployed for years and are used daily by billions of people around the world. There are several reasons for this lack of knowledge. Preferences are difficult to quantify holistically because they rely on and are shaped by numerous complex factors within a human being’s own psychology and external environment [Jameson et al., 2022]. Manipulating the preferences of real people experimentally comes with ethical challenges and often recommender systems are only considered in relation to the *behavior* of users and preference change is not acknowledged [Ashton and Franklin, 2022].

The unaddressed question of how recommender systems should handle preference change constitutes a problem for recommender systems research and deployment which we will refer to as the "preference change problem". Researchers in the recommender system community have begun to recognize this gap and work to establish a coordinated, multidisciplinary investigation into "preference science" [Franklin et al., 2022]. Additionally, simulation is being explored as a methodology which can study preference change without direct manipulation of real human preferences and in causally investigating the dynamics of complex socio-technical systems [Winecoff et al., 2021] [Lucherini et al., 2021].

In this thesis, we contribute a researched approach to the preference change problem which proactively promotes user agency. With any AI system, it is important for people to have agency and be "in-the-loop". This can help incorporate relevant contextual information which may be unaccounted for by the system designers and avoid unintended consequences [Stray et al., 2021]. Often, recommender systems will include users in-the-loop by inferring preferences only from behavior, and do not acknowledge cases where a user says one thing and does another [Ekstrand and Willemsen, 2016]. This can lead to a number of problems in recommender systems where recommendations are given in order to cater to a user's behavior, but leads to preference change which is not aligned with a user's long-term goals. Even recommender systems which use non-behavioral information like user profile attributes or item features may still not be reflective of these goals, as they often only ask "what do you want?" rather than "what do you want to want?". Franklin et al. [2022] recommends that systems bridge this conceptual gap by incorporating user "meta-preferences"; preferences about one's own preferences. These can be aligned with a user's preference change goals and express how conflicting preferences should be resolved within the recommender system. A system which incorporates meta-preferences has never been implemented, and the effects of such a mechanism remain unknown.

In this thesis we explore the following research question:

> How can simulations of recommender systems incorporate users' meta-preferences to result in user-directed preference change?

Our experiments make use of the new Tools for RECommender system Simulation (T-RECS). for the implementation and experimentation of a novel mechanism to incorporate user meta-preferences into content-based recommendation models. The tool was introduced by Lucherini et al. [2021] and provides a useful platform where simulation studies can be more easily reviewed and replicated. We take the code and dynamics of our simulation from a study on algorithmic confounding by Chaney et al. [2018] which was replicated in T-RECS by Lucherini et al. [2021]. The repository for these experiments can be found in the *T-RECS Experiments Github Repository*.

Our research takes the form of a simulation study with two distinct phases. We first create a simple model of content-based meta-preferences M in T-RECS and construct a novel feedback mechanism so they can be incorporated into a simulated content-based filtering algorithm. This extended content model is tested against a basic content model with no meta-preference mechanism as well as baseline ideal, random, and popularity models. The second phase of research consists of simulating and measuring the effects of preference change. There are no existing standards for simulating preference change in recommender systems research [Chaney, 2021], so we used the default mechanism in T-RECS and reveal opportunities for future work.

We organize our research into the following sub-research questions.

- RQ0: Are the results by Lucherini et al. [2021] about user and creator homogenization reproducible?
- RQ1: How can user meta-preferences M be simulated?
- RQ2: How can content-based recommendation models incorporate user meta-preferences M when determining predicted user preferences $\hat{U}$?
- RQ3: How can we simulate preference change among users?
- RQ4: Do recommendation models which incorporate user meta-preferences lead to preference change aligned with their target preferences?
- RQ5: What are the effects of user-directed preference change on recommendations?

The structure of the thesis is as follows. In chapter 2 we provide a background on the key technical information about recommender systems, our simulation methodology, as well as survey relevant literature on preferences and value-alignment. In chapter 3 we answer our sub-research questions and present the relevant methods, results, and discussion individually in a distinct section. Chapter 4 concludes the thesis by summarizing the results and answering the research questions. We include an appendix of our results with a log-log scale in appendix A.

Chapter 2

Background and Related Work

This chapter synthesizes preliminaries and related work and builds a case for user-directed preference change as a solution to the preference change problem. To begin, in section 2.1 we provide some relevant information about recommender system models, feedback mechanisms, and simulation techniques needed to understand the research in chapter 3. In section 2.2 we explore the preference change problem in detail, exploring its positive and negative outcomes, and critique the way preferences are currently elicited and represented by recommender systems. In section 2.3 we outline why and how the problem must be solved, presenting both theoretical and practical recommendations.

2.1 Recommender Systems

In this thesis, we take inspiration from [Jameson et al., 2015] and conceive of recommender systems as tools for helping people to make better choices.

Practically, recommender systems are technological algorithms implemented in digital environments which are used to narrow down some large catalog of items. They are widely used in domains such as social media algorithms, music, film, and book recommendations, as well as for shopping and commerce. Recommender systems have demonstrated their usefulness as online content catalogs grow to sizes which are unmanageable to sort through manually. To narrow down these catalogs and find relevant items, these systems use some model of recommendation to decide which items to recommend, and gradually learn and improve over time on data produced by the user as feedback. Recommender systems are often studied using "offline" methods, which use historical datasets and metrics like accuracy which typically assess a system's ability to successfully predict user behavior [Ricci et al., 2022]. Simulation is another methodology which is useful for exploring the larger dynamics and impacts of recommender systems on the larger socio-technical systems they operate within [Dong, 2022].

We present an overview of some models of recommendation which are relevant to this thesis in section 2.1.1. In section 2.1.2, we discuss explicit vs. implicit modes of feedback, and note how our research establishes another kind of mechanism. We defend simulation as our choice of methodology in section 2.1.3.

2.1.1 Models

There are a variety of ways recommender systems can consider items, and a variety of ways those techniques can be combined to generate recommendations. Recommendations can be made on the basis of content, the overlapping tastes of users, popularity, and many extensions combinations of these techniques. We refer to these distinct techniques for generating recommendations as “models” of recommendation.

In this thesis we specifically focus on recommendations that arise from the evaluation of the content of the items being recommended. This is because we conceptualize user-directed preference change to be prompted by the *content* of recommendations, rather than other factors. Models which use this approach are called content-based filtering or content filtering. They succeed by aligning a prediction of the kinds of “attributes” a user prefers with a model of what a certain item has. In the simplest case, a movie recommender that knows that a user has a history with a particular director might recommend a movie by that director which they have not yet interacted with. In this case, the director would be an attribute of the item, and the recommendation would be made on the basis of that attribute.

The attributes of an item differ depending on the domain, and can range from manually tagging important features (director, release year, actors) to a more semantically rich representation of the item which must be learned [Chaney et al., 2018]. For domains like social media where items are created at a higher volume than can be manually reviewed, content-based recommendations may rely on machine learning approaches to extract the semantic content of an item and generate attributes that way [Ricci et al., 2022].

In the original paper by Chaney et al. [2018], a simple content filtering model is simulated by directly matching user preference attributes U with item attributes I. To do this, the model keeps a model of *predicted* user preferences $\hat{U}$ which is updated each time the users chooses an item. In this way, the model continually learns the preferences of the user based on the user’s behavior. This model of content filtering is very simplified, and does not account for many nuances in the way a real-world recommender system might such as temporal relevance or user-item attribute mismatch, but it is effective in simulating the key dynamics of content-based filtering.

Other models or recommendations which are important to this thesis are popularity and random models. These are simple, non-personalized approaches to recommendation, and recommend every user either the most popular items in the former case, or a random set of items in the latter Ricci et al. [2022]. The simulation differs from Chaney et al. [2018] and Lucherini et al. [2021] in that it does not include the social filtering and matrix factorization models. These models did not have internal representations which were conducive to experimenting with facilitating user-directed preference shift and were excluded for purposes of simplicity.

2.1.2 Feedback Mechanisms

In recommender systems, there is a distinction between explicit and implicit feedback given to the model by the user. Explicit feedback often comes in the form of ratings (5-star ratings) and has fallen out of dominant use in state-of-the-art systems. Implicit feedback is now preferred due to its abundance, and tracks various aspects of the user's behavior when using the system (clicks, view time, retention) [Ekstrand and Willemsen, 2016].

While the increased use of implicit feedback led to advancements in the quality of recommendations by most metrics, Ekstrand and Willemsen [2016] makes the case that behaviorism is not enough—explicit feedback is needed to remedy cases where a user's behavior is not reflective of their thoughtful judgement. Additionally, there are some scenarios where the control provided by explicit feedback improves the quality of recommendations [Ekstrand and Willemsen, 2016]. The work by Wang et al. [2022] on User-Controllable Recommender Systems (UCRS) is relevant to this problem and is the closest existing work to this thesis. A key distinction is that UCRS are built to specifically avoid and mitigate the influence of filter bubbles and use offline datasets for evaluation, whereas this thesis is interested in recommender systems for preference change and uses simulation.

This thesis will explore a mode of feedback which is explicit in the sense that it is given by the user and not inferred from their behavior, but is generalized on the *kind* of items the user wishes to be recommended, rather than judgements on individual items. This is explored in detail with sub-research question one in section 3.2.

2.1.3 Simulation

Simulation is a useful research methodology for studying recommender systems because it combines empirical observations with theoretical insights [Winecoff et al., 2021]. It is particularly suited for theory development regarding systems with complex phenomena and non-linear effects [Dong, 2022]. By this we mean that the interaction of many the different components within a recommender system (users, items, creators, models) give rise to effects which are difficult to discern the cause of statistically. Qualitative methods are better at considering the system holistically but lack generalizability and precision.

Simulation is useful for this thesis specifically because it allows us to quantitatively evaluate preference change of individual users under different conditions, which would not be possible using an offline evaluation dataset. It also avoids the ethical issues which would inevitably arise when studying preference change on human subjects. However, it is important to note that simulation is not ethically neutral, it lends itself to a high degree of personal judgement on the part of the researcher in what effects are worth studying and how they should be simulated. While user-centric evaluation approaches are very important to incorporate the input of users into the design process, they are difficult when studying preference change because of measurement constraints and confounds such as the observer effect [Ashton and Franklin, 2022]. Franklin et al. [2022] call for a multidisciplinary effort to better understand preference change and recommender systems in order to provide more of an empirical foundation for research, but until such research has been performed simulation remains the best tool available.

Limitations

The central limitation of simulation as a methodology lies in the lack of established norms for studying recommender system and the corresponding heterogeneity between studies. Winecoff et al. [2021] outline the difficulty in synthesizing findings when studies will use different user and item representations, interaction dynamics, and metrics. Lucherini et al. [2021] discusses the difficulties in replicating simulation research, as code is often not shared publicly and even seemingly comprehensive descriptions leave room for interpretation. As solutions to this issue, Chaney [2021] proposes the adoption of vetted and critiqued models from economics as a solution to the heterogeneity in user choice representations. Additionally, a forthcoming paper aims to establish some norms within the recommender system simulation research community [Ekstrand et al., 2021].

T-RECS: Tools for Recommender System Simulation

The problem of simulation standardization was partially answered with the introduction of the flexible and open-source Tools for Recommender system Simulation (T-RECS) by Lucherini et al. [2021]. T-RECS provides a useful platform upon which these standards can be implemented and whereby simulation studies can be more easily reviewed and replicated.

We use T-RECS in this thesis due to its open source nature and flexibility. Many other simulation systems are focused on filter bubbles [Aridor et al., 2020] [Jiang et al., 2019] [Geschke et al., 2019] and polarization [Perra and Rocha, 2019]. It is expressive enough to study various kinds of recommender systems and other sociotechnical systems. We also benefited from the fact that the code for the replicated simulation of Chaney et al. [2018] was already publicly available saved a great deal of work in the development of this thesis.

2.2 Preferences and Behavior

If recommender systems are tools to help people make better choices, the recommender system's relationship to preferences that inform those choices is paramount to understand.

As we will discuss, behavior does not only arise from an underlying set of preferences, preferences will also arise from an environment mediated by behavior. This reciprocity is very visible in recommender system feedback loops, where the content shown to the user can affect their preferences, which then influence the recommender system. Despite this reciprocal relationship, recommender systems often conceive of user preferences as something to be predicted and fulfilled. In this, offline methods especially fail to consider the inherent dynamicism and impressionability of user preferences, and therefore remain blind to induced preference change. This preference change blindness leads to an incomplete understanding of the impacts of such systems on users and how exactly to handle recommender-induced preference change remains an open problem [Ashton and Franklin, 2022].

In this section, we review existing research on preferences with the aim to establish a solid theoretical and philosophical foundation for user-directed preference change. In section 2.2.1, we outline the current ways preferences are represented in recommender systems and the assumptions and flaws in this representation. In section 2.2.2 we discuss the issue of preference inconsistency and what consequences it has for recommender systems—as well as how preferences could alternatively be represented. Lastly, in section 2.2.3 we cover various facets of preference change in recommender systems, including how to distinguish between desirable and undesirable kinds of preference change, the positive and negative outcomes of recommender system-induced preference change, and the actual effect of recommender systems on user preferences.

2.2.1 Preference Representations in Recommender System Simulations

In the paper which presents many of the interaction dynamics used in this thesis, Chaney et al. [2018] use a formalization of user preferences as a combination of known and unknown utility.

A user may have certain preconceptions about what they like which influence their choices within a top-N recommender system, but the actual value of the item they choose is not known to the user prior to interaction. Chaney et al. declare the objective of the simulated recommender systems to estimate the total utility—to know both what the user prefers as well as predict the unknown added value an item will bring. Environmental factors like an attention mechanism also influence user choice, but the underlying preferences form the standard basis of the utility calculation function. The fixed underlying preferences are simulated as a discrete distribution over content-indicative attributes, and generated by a Dirichlet distribution which is parameterized in such a way as to simulate "long-tail" popularity among attributes. This representation was replicated in T-RECS by Lucherini et al. [2021], where the actual user preference U are represented as a discrete vector, where each value is indicative of the user's "preference" for items with that attribute. This preference is relative to all other attributes, since the user must make a choice at every timestep regardless of whether the presented items are a good match for the user.

Assumptions

There exist two assumptions in this formalization of preferences which particularly contradict with existing research into preferences and human decision-making. Firstly, this formalization simulates what Jameson et al. [2022] calls attribute-based choice, which is only one of many patterns of choice people may use when selecting items from a recommender system. Other choice patterns include socially based choice, policy-based choice, and experience-based choice [Jameson et al., 2022]. We continue using an attribute-based pattern of choice in this thesis, but further research should investigate how to simulate various choice patterns and what the consequences are of not doing so. The second assumption is present in the fixed nature of the simulated user preferences, the simulation does not acknowledge the possible influence the recommender system has over user preferences. This is the central assumption that we will challenge and attempt to address with this thesis, the next sub section will explain the research on preference consistency, and the following subsection will detail preference change more broadly.

2.2.2 Stability and Consistency

Many recommender systems use preference models which are stable (unchanging) and consistent (the same choice is made from the same input), despite there being research that challenges this.

Research from psychology and behavioral economics demonstrate that human decision making is inconsistent [Franklin et al., 2022]. This behavioral inconsistency can arise from factors such as mood, whether the decision is made privately or in a context with social norms and expectations, and conflicts between a person's short and long-term preferences [Ashton and

Franklin, 2022]. Research has shown that a decision-maker's preferences can be learned in the course of making a decision, such as learning what attributes are available and therefore which to value [Dzyabura and Hauser, 2019].

Interestingly, this lack of stability partially led to the downfall of pure-accuracy metrics like the root-mean squared error (RMSE) as a comprehensive measure of success for recommender systems. Designers found that using historical data to predict future choices would always fall short due to the instability of user preferences. Without the existence of strongly-held preferences to accurately predict, alternative metrics for diversity, serendipity, and eventually engagement began to come into play [Seaver, 2019].

Preferences and Identity

Psychological research has also shown that people will accept preferences assigned to them, even if the assigned preferences contradict past indications [Hall et al., 2012]. This relates to the notion that categories are integral to one's experience of personal identity. The construction of identity occurs partially in the experience of accepting and rejecting certain categories [De Vries, 2010]. This connection means that the notion of user preferences as conceived by recommender systems are important to the construction of the user's identity, and that identity—similarly to preferences—is fluid and subject to influence. This motivates recommender-induced preference change as very important to address. If recommender systems are influencing the behavior, and therefore also the preferences, and to some degree the identity of their users, how can these systems be designed to steward that responsibility safely and respectfully?

Representing Dynamic Preferences

There is research beginning to be done on studying user preferences as dynamic. Sanna Passino et al. [2021] studied the success of a Long-Term Value (LTV) reinforcement learning recommender system which utilized a dynamic representation of user preferences called the Preference Transition Model (PTM) which an anticipate probable trajectories of preference shift. This is a promising start, but the research did not explicitly address *how* preference change should be handled. If various trajectories are probable, how should the recommender system proceed? How should the trajectories be evaluated and prioritized?

Ashton and Franklin [2022] conclude their review by saying that there is no singular or comprehensive answer to the preference change problem, but in any solution it seems necessary to provide the system with an indication of *which* set of simplified preferences it should model and reflect back to the user. This sentiment is explored further by Franklin et al. [2022], who defines “meta-preferences” as user’s preferences about their own preferences and “preference change preferences” as a user’s wishes for how and why their preferences should change.

One approach to the preference change problem is to make the incorporation of these a point of autonomy. Recommender systems can leave it up to the user to specify which preferences they wish to develop and explore with the system. This avoids the need for the designers to determine what specifically constitutes good or bad preference shift. A case is made for incorporating the users in a “bottom-up” approach in section 2.3.2 and the merits of top-down vs. bottom-up approaches is discussed 2.3.3.

2.2.3 Preference Change

The notion of dynamic preferences leads us to consider a central concept of this thesis, which is the notion of preference change—a measurable change of a person’s preferences in some regard. The exact nature of what preference change means depends on the domain, in regards to entertainment media this might mean a shift in the kinds of genres a person engages with, but it can also mean more fundamental shifts such as a person’s future goals or beliefs about the world.

In their work in modeling dynamic user preferences, Sanna Passino et al. [2021] measure preference shift of a sample of Spotify users by computing the total variation distance $\frac{|U_1 - U_t|}{2}$ between a user’s preferences (represented as a distribution of the genres they listened to in some window of time) from the beginning of the collected data U_1 to later time periods U_t. They observed users shifting away from genres they listened to. Similarly, in their work in modeling degenerate feedback loops in recommender systems, Jiang et al. [2019] measure preference shift for a finite item set with the L_2 norm $||U_0 - U_t||_2$. With simulation, they uncovered some factors which contributed to the speed of ”degeneracy” or induced preference change. A similar measure to Jiang et al. [2019] is used in our research in measuring simulated preference change in section 3.4.3.

Natural vs. Induced Preference Change

Recommender systems are not the only cause of preference change—user preferences are also going to evolve as a result of other external and internal factors. Given this, both Carroll et al. [2021] and Ashton and Franklin [2022] distinguish between “natural” preference change vs. preference change induced by the recommender system.

To delineate the difference, we must first understand the mechanisms by which the recommender system can affect user preferences. Cognitive biases such as the mere-exposure effect, availability bias, and anchoring can cause our preferences to adapt to things we are familiar with [Ashton and Franklin, 2022]. For example, if a user of a movie streaming platform is recommended mostly action movies since the system has learned they are particularly likely to engage that genre, over time, a user could grow to prefer it for reasons of familiarity, ease, and eventually nostalgia. It is also important to note that certain types of content is more likely to cause preference change than others. Politically-charged and conspiracy content is especially effective when it discredits other sources of information as unreliable and ensures users become reliant on one narrow band of trusted sources [Van der Linden, 2015].

In their simulation investigating the prevention of induced preference shifts Carroll et al. [2021] defined "natural" preference shifts as what the user would experience if interacting with the content with full information and no recommender system. They define a "safe zone" which allows for dynamic preferences but penalizes manipulation of preferences to make users more predictable, which is a recognized problem for systems which optimize for long-term engagement. This is possible to measure in simulation studies (equivalent to the "ideal" recommender in T-RECS), but is difficult to perform in real-world settings. The best a system can do is try to predict when a policy might cause an unnecessarily strong induced preference change and prevent it [Carroll et al., 2021]. This may be effective in avoiding some of the worst effects of induced preference change (discussed in the next sub-section) but beyond reactive prevention, there is a need to explore the possibility of using the recommender system for *enhanced* preference change aligned with a user's goals and meta-preferences.

Negative Outcomes of Induced Preference Change

We give three examples of concerning impacts of recommender system-induced preference change.

The first is change toward insulated preferences. A lack of exposure to different viewpoints as a result of homogenized recommendations leads to users becoming stuck in "filter-bubbles" [Geschke et al., 2019]. Filter bubbles have important political implications in news feeds (facts and sensemaking), social media (sensemaking and debate), and other systems recommending ideologically-laden content. Filter bubbles are what allow for online "echo chambers" to form, where content which aligns with certain ideologies is recommended to people who already agree with it and the ideology is thus reinforced. Geschke et al. [2019] notes that echo chambers are prone to develop rhetoric which is more radicalized and polarized. The power of psychological inclinations to conform with group social norms combined with a highly skewed understanding of alternative viewpoints can lead to rapid escalation and demonization of "the other side". These processes can lead to the degradation of democratic processes of debate and heighten the potential for violence and terrorism connected to radicalized online groups

[Geschke et al., 2019].

Another worrying case is that of preference change caused by active manipulation by a small group of users. Most recommender systems which rely on user activity for popularity signals are vulnerable to strategic manipulation by third-party actors [Chakraborty et al., 2019]. As an example, the demographics of the most active users social media sites such as twitter are not aligned with the demographics of the overall population. Not only would this demographic bias have passive effects, but the imbalance can be intentionally exploited through the use of automated bots to further tip the composition of the crowd and their opinions. This manipulation can have very damaging societal effects, especially in light of the promotion of political propaganda in combination with digital echo chambers, which can amplify the effects of the propaganda. This means that a few targeted bots spreading targeted content can lead to widespread influence on the opinions of real users [Chakraborty et al., 2019].

Our final example is when recommender systems facilitate a shift in preferences toward addictive content. This tendency is especially present in systems which optimize for engagement, or what Seaver [2019] calls "captivation metrics". These metrics prioritize the prolonged and profitable use of the recommender system by users, and take a user's implicit behavior as a measure of value. Systems which use them tend to serve content which facilitates repeated engagement, since compounding repeat consumption ensures the growing long-term value of the system [Seaver, 2019]. Hasan et al. [2018] found that, among other factors such as a lack of self-control and self-esteem, the use of recommender systems lead to excessive video streaming. Andreassen [2015] give a comprehensive review of social media site addiction as a psychological phenomenon, but the specific role of recommender systems are not discussed. There does not seem to be a comprehensive study or review of recommender systems and addictive content, and given the presence and danger of this phenomenon, a researched multidisciplinary investigation is greatly needed.

Developers of recommender systems may not think of what they are optimizing for as addiction, but rather "satisfaction". Seaver [2019] notes that considering a user who is highly engaged with a recommender system to be satisfied allows developers to mediate the different goals of the people building the system and the people using it. Despite developers' good intentions, this conflation of engagement and real satisfaction is a fiction, with the most extreme and obvious conflict in the case of addictive content. In this case, users are extremely engaged with the recommender system but find themselves unsatisfied with their behavior in retrospect. Addiction is a clear example, but Seaver [2019] asserts that the spectrum between freedom and coercion is not clear-cut, and discards this binary distinction. Instead, the author notes that engagement-optimizing recommender systems work more to exploit, rather than deny, the autonomous agency of users. As Milano et al. [2020] notes, the question is not how to avoid the captivating tendencies of recommender systems, but rather how to use them in such as way as to holistically benefit the users. The next section will discuss this possibility.

Positive Outcomes of Induced Preference Change

Self-development lies not only personal processes reflection and growth, but also in recognizing and changing the systems outside the self which influence it.

Not all preference change is undesirable, and there is significant potential for using recommender systems to promote self-development through the consumption of developmentally valuable items. Recommender systems can do more than just minimize filter bubbles, they can also be designed to broaden our perspective of the world beyond what can be encountered with manual exploration. Work is starting to be done on recommender systems for developing new preferences and goals [Liang, 2019] and Knijnenburg et al. [2016] explores recommender systems for self-actualization.

If preferences can be supported to change in certain "self-actualizing" directions, how can we say what these directions are and how they are different from other avenues of preference change? Lades and Delaney [2022] presents a an ethical framework for defining what constitutes good or bad behavior change for policy, and the theory could be applied to design choices for recommender systems. Elster [2016] presents a more philosophically motivated theory, delineating preference change caused by learning and experience as more desirable.

For this thesis, we leave the decision of deciding what is good preference change up to the users, therefore avoiding the need for this complicated philosophical problem to be quantified within the optimization criteria of recommender systems. This is still not a perfect solution, since leaving it completely up to the individual may ignore important considerations about societal well-being and social cohesion.

2.3 Value Alignment

If recommender systems are to avoid potentially harmful and unintended consequences, they must be aligned with more than just the values of the system designers.

Value alignment is a central theme in conversations regarding the impacts and ethics of human-made systems. As a concept, it involves the examining of values embedded in technology and questioning their alignment with the values of the people they effect in direct and indirect ways. By considering value alignment, we consider the supposed purpose of recommender systems, what they are actually used for, and postulate what they could or should be used for.

Section 2.3.1 will present three challenges to value alignment in recommender systems. In section 2.3.2 we discuss some theoretical approaches to addressing these challenges. Finally, in section 2.3.3 we illustrate the place of this thesis within the theory and explain why current approaches fail to align to the end goal we envision for recommender systems.

2.3.1 Three Challenges

This thesis identifies three relevant challenges to designing well-aligned recommender systems. The examples presented in this section do not constitute a full inventory of the challenges, for a more detailed overview of some documented ethical challenges of recommender systems, see the paper by Milano et al. [2020].

1. There is a lack of meaningful control

A common requirement for ethical AI is the need for a human to be "in-the-loop" of the AI system. This means that decision-making is not simply relinquished to the AI, but is made with a human who has meaningful control in the process of the decision [Jameson et al., 2015]. Recommender systems may at first seem exempt from this requirement since ultimately it is the users who chose which item to consume, but if the user has little to no explicit control over *how* the system narrows down choices, are they really in the loop to the best degree possible? As we discussed in section 2.1.2, sometimes user behavior will misalign with their meta-preferences. In this case, the user has no convenient mechanism to exert meaningful control over their recommendations.

The question of meaningful control is also blurred in the use of recommender systems as persuasive technologies. Seaver [2019] discusses the emergence of persuasive technologies and the field of "captology" (originally invented to express the field of "computers as persuasive technology" and later linked with theories of traps and capture). Seaver [2019] explains that captology as a field and practice is interested in technology which can change the beliefs and behaviors of people. Such technology uses behaviorist theories which understand people as "habitual minds with tendencies and compulsions that make them susceptible to persuasion and targets for capture." The principles of persuasion employed by captological systems have been critiqued for being coercive rather than persuasive—the key difference being the amount of freedom one has in the decision and the symmetry of information. You will always know when someone is trying to persuade you, but coercion is most effective when it is below conscious awareness. Users of recommender systems can hardly be said to be consciously aware of all the techniques the system is using to nudge them toward maximum engagement and profitability, and this constitutes another lack of meaningful control.

2. There are sometimes cases where past data is insufficient and short-term wins are favored over long-term satisfaction

While we are indeed creatures of habit, we are not deterministic machines. Recommender systems should not maintain the assumption that past data is sufficient to predict and shape future preferences. Users of social media often generate a large amount of engagement when emotionally outraged [Rathje et al., 2021], but may not find these interactions to be a valuable use of their time retrospectively. Stray et al. [2021] note that retrospective evaluation is of a higher quality than immediate implicit feedback because it can distinguish between behavior and the more informed, deliberative judgement of users.

This challenge is a recognized problem for AI metrics in general, where optimizing for narrow metrics can lack long-term success due to the inability for these metrics to measure and account for the greater complexity of the system [Thomas and Uminsky, 2022].

3. Asymmetric Design Power

Even if users recognized the disconnect between behavior and meta-preferences, and saw the need for a mechanism beyond implicit feedback the users have no real way to influence the design of the system to better fit their needs.

Users also are often unwittingly subject to various subversive forms of influence when interacting with a recommender system [Seaver, 2019]. The designers hold asymmetric power in their ability to make design choices which benefit their own interests and in choosing which problems get recognized and addressed. This asymmetry of power often leads to externalities (such as the environmental consequences of amplifying consumption of unsustainable products) to be pushed to users and non-users in such a way that profit can still be optimized, at least in the short-term [Stray et al., 2021].

2.3.2 Theoretical Approaches

To address these challenges, the theoretical framework of value-sensitive design provides a useful tool to aid in building aligned recommender systems.

Many of the challenges above relate to Ekstrand and Willemsen's [2016] assertion that behaviorism is not enough. Users must be listened to in a more thoughtful manner than simply making predictions based on their behavior. Stray et al. [2021] assert that behaviorist assumptions can be over-come in recommender systems if they are designed around a user's informed, deliberative judgement and interactively learn the values of users.

Value Sensitive Design (VSD) is a theory of design concerned with the consideration of ethics and the values embedded in technology [Friedman et al., 2013]. We take values to mean "what people or a group of people think is important in life" [Friedman et al., 2013]. Ekstrand and Willemsen [2016] articulates the place and importance of values in their call for recommender systems to listen to users.

> Values are present in all system designs. The question isn't whether a system embeds some set of values; rather, whose values (and what values) does it embed? Are these values explicitly articulated and subject to discussion? Are designers transparent about the reasons for decisions? Giving credit to user perspectives is itself a value that can be included in or rejected by a technological process, and including user input can enable debate and discussion about the other values the system embodies [Ekstrand and Willemsen, 2016].

The strength of VSD lies in that it makes these embedded values explicit. VSD opens a conversation about trade-offs and consequences as opposed to simply accepting certain features like bias for short-term gains or large power asymmetries as inherent and inevitable.

VSD provides a useful framework to think about value-alignment on a high-level, but ethical frameworks do not easily translate to actionable design choices [Mittelstadt, 2019]. A principle like "give the user meaningful control" can be interpreted any number of ways. The next section will discuss the goals and values we envision for our research and practical challenges to alignment with those goals.

2.3.3 In Practice

This thesis will simulate recommender systems which continually incorporate user meta-preferences via a novel feedback mechanism. In this, we investigate the success of this mechanism to solve the lack of meaningful control and misalignment with long-term user goals by giving users more explicit power over their recommendations. In contributing this research, we envision a future where recommender systems are designed to actively enhance societal well-being and individual self-actualization. Though this is an admirable goal which many designers may agree with, it is important to address *why* current approaches to recommender system design are not aligned with this goal.

Ekstrand and Willemsen [2016] note that we can glean the values of recommender systems by looking at their optimization criteria and evaluation metrics. Right now, most large-scale commercial recommender systems optimize for metrics designed for commercial success in the short-term, but lead to second-order effects like addiction, radicalization, and polarization. Despite this clearly being a problem and industry being called on to stop these harms, the solutions implemented are usually reactive, and do little to change the optimization criteria or the corresponding values embedded within them.

Mittelstadt [2019] acknowledges that there are real commercial reasons for this. The merits of VSD are recognized in academia but are not necessarily going to be prioritized in commercial contexts. If a company were to invest the time and effort of building a value-sensitive recommender system which was which was less commercially successful than a less thoughtfully designed one, they would experience a short-term loss in time and resources and competitively under-perform. Technology companies are generally incentivized to code first and address possible consequences later [Schultz, 2019].

Despite the limited practical applications, we believe it is intensely important to align recommender systems research in service of a meaningful goal. Incentives and optimization criteria come and go, and recommender system research must move beyond the myopic drive toward marginal gains. We hope this thesis can contribute to the foundation of research needed for designing recommender systems which can improve both the people and societies which use them.

Chapter 3

Research

If behaviorism is not enough [Ekstrand and Willemsen, 2016], and user preference change is an unaddressed problem in recommender system research [Ashton and Franklin, 2022], users should be given a way to gain agency over their recommendations and therefore also their preferences.

Due to the ethical problems of performing user studies which seek to actively change user preferences [Ashton and Franklin, 2022], we will use T-RECS to simulate a content-based recommender system which incorporates user meta-preferences into the process of recommendation and subsequently measure the affected preferences of simulated users. A visualization of the altered T-RECS simulation can be seen in figure 3.1.

In section 3.1, we replicate the original simulation by Lucherini et al. [2021] and explain the dynamics of our simulation in detail. Section 3.2 will detail the expansion of the user representation to include meta-preferences and section 3.3 will discuss the process of incorporating our meta-preference mechanism into the model. In section 3.4, we devise a way to simulate and measure preference change. After all the necessary expansions to the simulation have been addressed, we present the effects of the meta-preference mechanism for user-directed preference change in section 3.5. Finally, section 3.6.1 surveys the larger system-wide effects of the meta-preference mechanism and discusses some of the societal implications of our work.

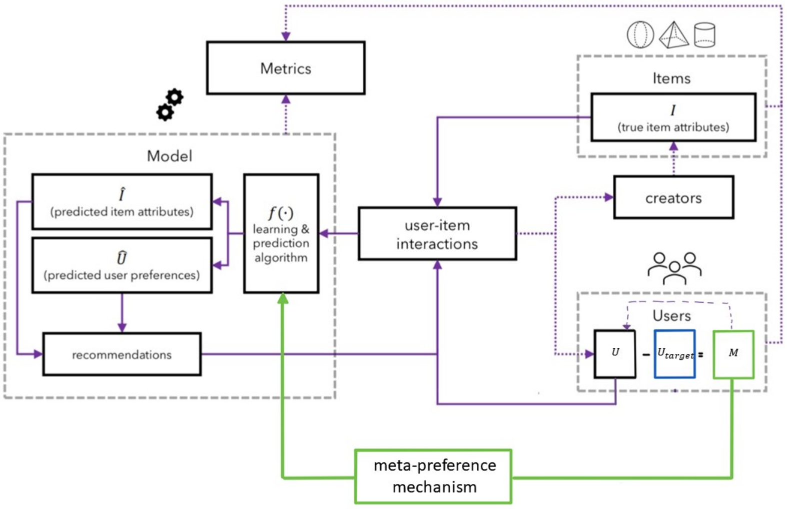

Metrics
Items
I (true item attributes)
Model
$\hat{I}$ (predicted item attributes)
$\hat{U}$ (predicted user preferences)
$f(\cdot)$ learning & prediction algorithm
user-item interactions
creators
Users
U
U_{target}
M
recommendations
meta-preference mechanism

Figure 3.1: An overview of the simulation as presented by Lucherini et al. [2021] with the proposed expansion of a meta-preference mechanism. The user representation is expanded to include the target preferences of the user, U_{target} and their meta-preferences M, the latter of which is passed to the model via a secondary interaction mechanism. The meta-preferences M are obtained by taking the difference between a user's actual preferences U and target preferences U_{target}. Grey dotted lines outline conceptual modules, solid arrows indicate concrete inputs, and dotted arrows indicate potential feedback mechanisms.

3.1 RQ0: Creator Homogenization Experiment Replication

To begin our work in simulating user-directed preference shift, we first replicate the creator homogenization experiments by Lucherini et al. [2021] in T-RECS and explain the relevant dynamics of the simulation environment.

The addition of dynamic content creators to T-RECS by Lucherini et al. was a constructive replication of the experiment by Chaney et al. [2018]. They define a constructive replication as an experiment which replicates and then expands upon some previous work, adding new scientific insights [Lucherini et al., 2021]. The addition of content creators into the dynamics of the simulation allowed Lucherini et al. to observe not only user homogenization, but also creator homogenization. Given that they had already adapted T-RECS to Chaney et al.'s simulation framework, they were both able to save time in determining the dynamics of the simulation, as well as allow for the direct comparison of the two simulations.

In this thesis, we also reuse the assumptions and simulation parameters of Chaney et al. and build upon the technical implementation by Lucherini et al.. Given the lack of standards in recommendation simulation research, building upon the established work will make the results of this thesis easier to evaluate and compare, as it will share many of the same strengths and limitations as previous work. Additionally, it means that the same metrics can be used to evaluate the effects of the proposed changes on between-user homogenization in section 3.6.1.

In this section we answer the following sub-research question:

> Are the results by Lucherini et al. [2021] about user and creator homogenization reproducible?

Thereby answering RQ0 and testing the functionality and computational feasibility of T-RECS before subsequent experiments. In section 3.1.1, an overview of the simulation will be given, explaining the user, item, and model representations, interaction dynamics, and measurement metrics used by Chaney et al. and Lucherini et al.. Then, the specific methodological details of the replication will be discussed in section 3.1.2. Finally, in section 3.1.3 the results of the replication are presented in comparison to the results of Lucherini et al., with no significant differences observed.

3.1.1 Simulation Overview

The simulation is implemented in T-RECS with the following characteristics.

User and Item Representation

Attributes A are core to Chaney et al.'s and Lucherini et al.'s simulations. The replicated simulation uses the default 20 attributes, $|A| = 20$ as specified by Chaney et al.. Items are represented by a vector of binary attributes, $A \in \{0, 1\}$. The users are represented by a vector U of length $|A| = 20$ representing a user's known preferences which correspond to the degree to which the user thinks they will enjoy an item with that attribute. Our simulation uses $N = 100$ users in total, matching the original simulations. Given that the users are not omnipotent, they only have partial knowledge of how much a recommended item will correspond to their preferences. In the simulation, this is expressed by a function of the utility of some item I for some user U. This utility is partially known to the user, and the user chooses to interact with one item per timestep based on this partially-known utility and an attention mechanism. After interacting with an item, that item is not recommended to the user again.

New Item Generation

In the simulation dynamics presented by Chaney et al. [2018], new items are generated at each timestep and randomly incorporated into users' recommendation lists. This random interleaving serves to simulate users engaging with items beyond their recommendations, but the process of new item generation is limited to a fixed distribution. In the expansion by Lucherini et al., the item catalog was adapted to be dynamic, with content creators generating new items based on their creator distribution C.[1] Mathematically, C is a vector of length A, and each entry represents the Bernoulli probability a creator will create an item with that attribute. The creators adapt this distribution to respond to user feedback in each timestep, shifting C toward the attributes of their most successful items.

The item catalog model of Chaney et al. is more appropriate for a movie or book recommendation platform, where the assumption of a mostly fixed catalog is reasonable and items are not so explicitly adapted to user feedback. The dynamic content creators of Lucherini et al. better simulates fast-moving content on social media platforms like YouTube or Instagram where user feedback and item success has a more immediate impact on creators.

Models of Recommendation

Another key piece of the simulation is the representation of the recommendation system models. Both Chaney et al. and Lucherini et al. used simulations of four common recommender system models: popularity, content filtering, matrix factorization, and social filtering, plus ideal and random models for baseline comparison. Details of these models can be found in 2.1.1. All models represent and process user preferences differently and generate unique predicted user preferences $\hat{U}$. The recommendations are then based on the predicted user scores $\hat{S}$ which are a function of the predicted user preferences $\hat{U}$ and predicted item attributes $\hat{I}$. Though it is possible to simulate recommendation models which must also learn the item attributes, both Chaney et al. and Lucherini et al. used binary attributes and perfect knowledge of item attributes $\hat{I} = I$ in the interest of simplicity. An overview of the simulation dynamics in Lucherini et al.'s content creators experiment can be seen in Figure 3.2.

Metrics

Finally, Lucherini et al. measured the entropy of a creator's distribution c and averaged that value across all creators to the average measure within-creator homogenization at each time-step. This metrics, the average creator

[1] The item catalogs of Chaney et al. and Lucherini et al. [2021] are both technically dynamic in that they change over the course of the simulation as new items are added. Lucherini et al. take a "fixed" catalog to mean that the items are generated from a fixed distribution, not that the catalog itself is unchanging. A "dynamic" catalog is then one in which items are generated from changing distributions. This thesis will use such terms similarly.

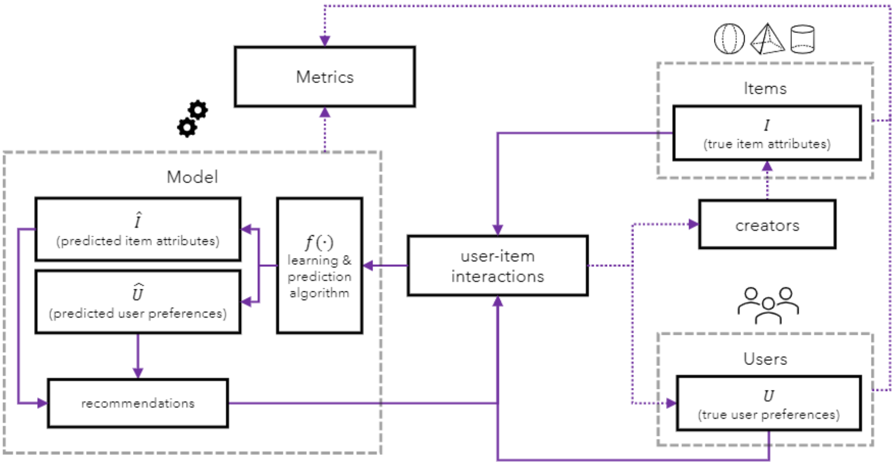


Figure 3.2: An overview of the simulation as presented by Lucherini et al. [2021]. Grey dotted lines outline conceptual modules, solid arrows indicate concrete inputs, and dotted arrows indicate potential feedback mechanisms.

entropy (ACE), essentially measures whether a creator uses a broad or narrow distribution when creating items for the system. In the preliminary investigation into the effects of dynamic content creators on the homogenization of user behavior, Lucherini et al. had to use a different metric than Chaney et al. due to that fact that the Jaccard similarity was not suitable for a dynamic item catalog. The new metric was the Average Pairwise Distance between Mean Consumed Items (APDMCI) which compares the means of the sets of consumed items for two users, and computes and averages that across all users.

3.1.2 Methods

The content creators experiment by Lucherini et al. [2021] was replicated using the source code in the *T-RECS Experiments Github Repository* on the 10th of March 2021.

The original simulation results presented in the paper ran 400 trials for each of the recommendation algorithms. These were were then averaged together to create the graphs demonstrating the average entropy of the creators' item-generating distributions as seen in Figure 3.5. Due to time and computational resource limitations, this replication used only 50 trials instead of the original 400, as recommended in the instructions for recreating the results. This did not prove to be a problem for the replication and the same trends of creator homogenization were observed with minimal difference from the results of Lucherini et al. [2021].

3.1.3 Results

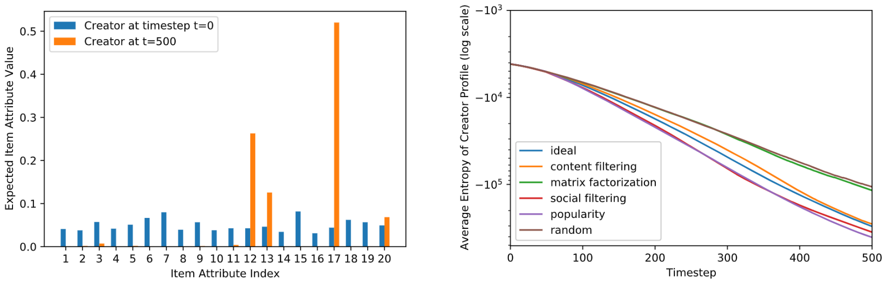


Figure 3.3: Original results of the content creators experiment by Lucherini et al. [2021] with 400 trials.

The results of the replicated simulation very closely match those of Lucherini et al.. In Figure 3.5, we can see that, as in Figure 3.3, creator homogenization occurs. As the simulation runs, the ACE for all recommender models decreases, and the rate of this decrease becomes more drastic over time (the scale of the axis of measurement in 3.5 is logarithmic).

This is visualized for one random creator in Figure 3.4. At the beginning of the simulation, the expected item attribute value is around the same for all attributes, meaning that items generated by a creator could have any of the attributes with an equal probability. This also means that at the beginning of the simulation the catalog of items generated by a creator are diverse in attributes when taken as a group, and the items individually are less similar to each other. By the end of the simulation, after the creator has adapted their output to what has been successful among users, the expected attribute value is concentrated around only a few attributes. This means that the items generated by the creator in later timesteps are more homogeneous in their attributes.

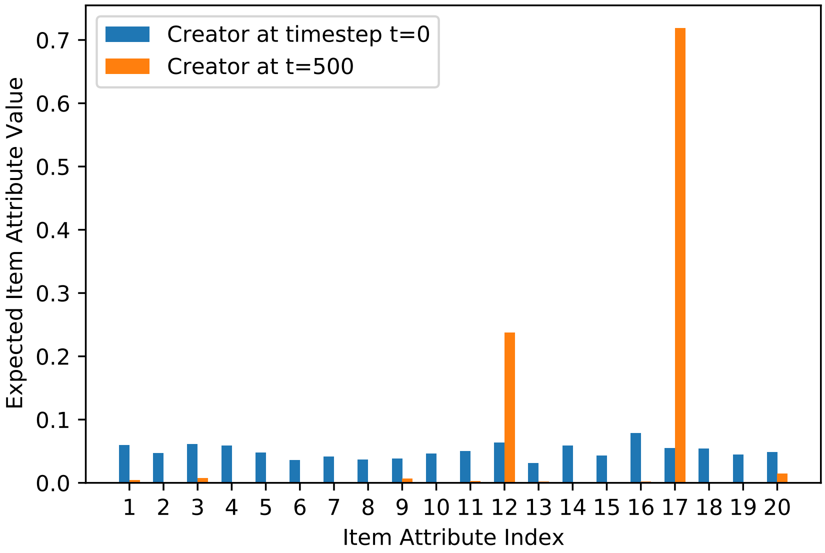


Figure 3.4: A random creator distribution at the beginning and end of the simulation. At timestep zero all attributes have an approximately uniform chance of being included in a generated item, but by the end of the simulation, the distribution has concentrated around two attributes. This means that new items generated by the creator will very likely only contain those one or two attributes.

The speed and degree of the homogenizing effects depends on the recommender system used. Both the results of this replication and that of Lucherini et al. show that the effects are greater for popularity-based recommendations and more moderate for random recommendations, but all models have a homogenizing effect on creator output. This is replicated in Figure 3.5 where the ACE is averaged and plotted for all models.

As mentioned by Lucherini et al., the metrics used in this simulation do not measure the overall homogenization of items so the catalog of items in the whole simulation could potentially still be diverse. Additionally, they stress that within-creator homogenization is not necessarily negative, as it could reflect creators settling into specialized niches and learning the interests of users.

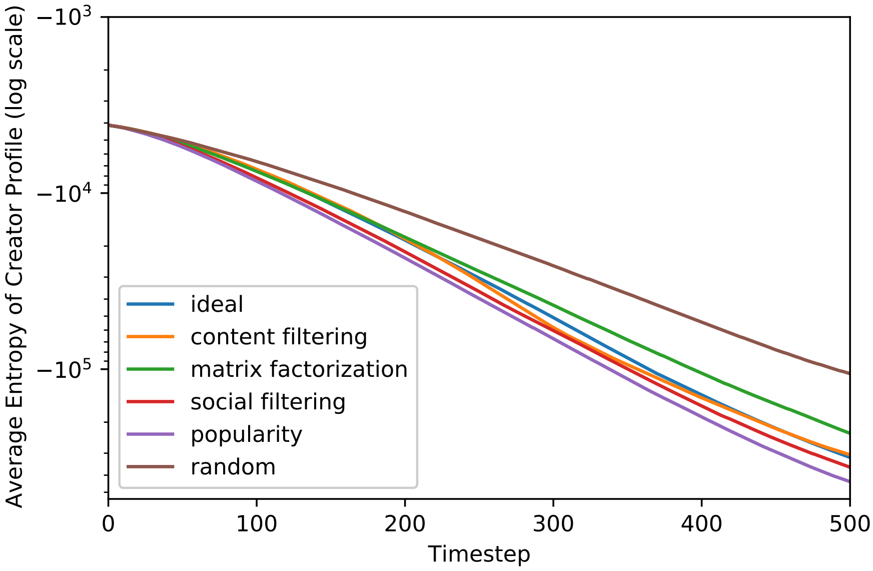


Figure 3.5: The ACE at each timestep over the course of the simulation. For all models, the average entropy decreases at an increasing rate.

3.2 RQ1: Representing Meta-Preferences

There has been increasing recognition of the need for further research into preferences and preference change if recommender systems are to be deployed responsibly. Franklin et al. [2022] define two aspects of preference which current recommender systems fail to account for, the first of these being "meta-preferences". These are preferences which users have about their own preferences. Additionally, "preference change preferences" are preferences the user has about how they would like their preferences to change in the future. Both of these concepts are very important for recommender systems to consider, since a blindness to these aspects of preference may lead to a recommender system which fails to put the user in-the-loop of their own preference change. Until now, there has been no research into the best way to represent user meta-preferences in a simulated environment, therefore, we had to create one ourselves. In this section we answer the following sub-research question:

> How can user meta-preferences M be simulated?

We present our representation of meta-preferences M in T-RECS as the difference between some target user preferences U_{target} and actual user preferences U and and discuss the assumptions and limitations of our approach.

3.2.1 Target Preferences

In our simulation, we expanded the user representation to include another distribution of preferences, which we call a user's target preferences U_{target}. This distribution differs from the user's actual preferences U, which are used in determining the behavior of users when interacting with the recommender system. Rather, U_{target} represent the distribution of preferences which the user *wishes* to have. This concept loosely corresponds with Franklin et al.'s 2022 idea of "preference change preferences".

This distribution is generated using the same techniques found in the user formalization by Chaney et al. [2018]. U_{target} is a $|U| \times |A|$ array where each value corresponds to the degree to which the user wishes to consume items with that attribute. The target preferences are drawn from a Dirichlet distribution which is parameterized by a global distribution of target attribute popularity. This means that some attributes are more popular than others as target attributes. An example of this might be a popular demand for health or educational content over more shallow short-term value items, as becoming healthy or more knowledgeable are popular long-term goals. The popularity parameters for U_{target} are different than the popularity parameters used by the actual user preferences U, and the tension between them is reflective of the observation that the long-term goals of users may not always align with their short-term behavior Ekstrand and Willemsen [2016].

In our formalization, U_{target} is independent from U, completely known to the user, and does not change over the course of the simulation. This is for ease of measurement and demonstration, but the limitations of such an approach are discussed below in section 3.2.3.

3.2.2 Meta-Preferences

In the expanded user representation, the users' meta-preferences M are expressed as the difference between the actual and target user preferences, $U_{target} - U$. This is visualized in figure 3.6. As the user's actual preferences U drift toward the attributes of the items they consume (our preference drift mechanism is discussed more in section 3.4), the meta-preferences are recomputed.

In the simulation, M is a $|U| \times |A|$ vector of values between -1 and 1, with values close to -1 being a strong meta-preference against that attribute (the user would like to consume less items with that attribute), values close to 0 being neutral (the user is happy with how much they consume items with that attribute), and values close to 1 being a strong meta-preference for that attribute (the user would like to consume more items with that attribute in the future).

M is computed and eventually passed to the model as a source of feedback, weighted by a strength indicator α. The exact mechanism of the incorporation of M into the model is discussed in section 3.3.

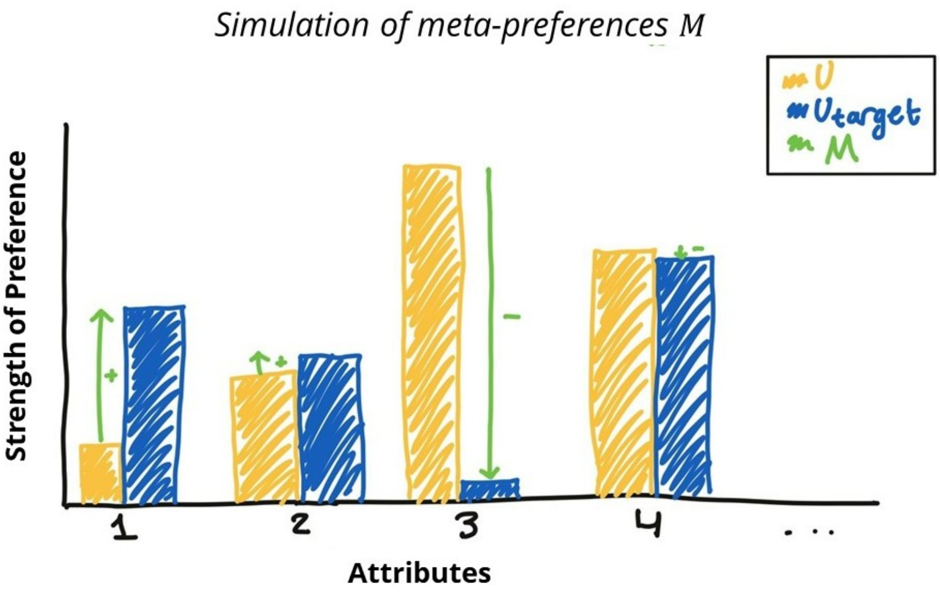


Figure 3.6: Visualization of how user meta-preferences are simulated as the difference between target and actual preferences. The final simulation uses a total of $|A| = 20$ attributes.

3.2.3 Assumptions and Limitations

The formulation of user meta-preferences presented in this thesis is simple and useful for demonstrating broad trends in the simulation but it has a limited degree of realism due to several factors.

Firstly, the discrete nature of the target preference distribution shares the same limitations as the preferences models by Chaney et al. [2018] and Lucherini et al. [2021] in that it only represents user preferences as attribute-based choice. This is only one of many patterns of choice, and reducing preferences down to strictly content ignores other factors like the experiential or social value of an item Jameson et al. [2015].

Another large limitation is the stationarity of the target preferences. Though we have modified the actual preferences to change throughout the course of the simulation (see section 3.4) the target preferences do not change. It is incorrect to assume that a user has some “ideal” distribution of preferences which are independent from their actual (behavioral) preferences, and it is reasonable to assume that, just as their short-term preferences are influenced by their environment, their long-term target preferences might change too. However, for the purposes of this simulation we do assume at least some short-term stability in target preferences, as we conceive them to come from an informed and deliberate long-term judgement, which is in theory a more consistent source of preferences than immediate choice [Stray et al., 2021].

Though we adopt the meta-preference terminology of Franklin et al. [2022], we do not distinguish between meta-preferences and preference change preferences in our simulation. We assume that if a user would like to consume less content with a certain attribute, that means they would like to shift their preferences away from that attribute. In practice this might not always be the case, but in the dynamics of our simulation they are conflated.

Finally, we present no concrete solution for exactly how meta-preferences may be evoked from the user, and we assume perfect, constant feedback with no noise or inconsistency. Assuming that all users will give constant feedback and diligently adjust their recommendations to better align with their target preferences is clearly unrealistic, and the system would favor those who have the time and technical knowledge to practice feed-curation. Still, we believe that including a well-design meta-preference mechanism would make curation easier and more accessible, especially because users who wish to control their recommendations today can only do so with imprecise item-by-item judgements and implicit cues. Finding an effective way to evoke user meta-preferences would be an interesting direction of further research and the work of Wang et al. [2022] and Bostandjiev et al. [2012] could be useful in this regard.

3.3 RQ2: Incorporating User Preference Judgements

In order to simulate user-directed preference shift, we must change the recommendation system to give recommendations which are more aligned with the kinds of content users would like their preferences to shift toward.

In this section we answer the following sub-research question:

> How can content-based recommendation models incorporate user meta-preferences M when determining predicted user preferences $\hat{U}$?

We first discuss how we adjusted the predicted user preferences in T-RECS in section 3.3.1. Then we motivate our custom metric of Predicted-Target Similarity in section 3.3.2 and present our results in section 3.3.3. We conclude this section with a discussion of our approach in section 3.3.4.

3.3.1 Adjusting the Predicted User Preferences

In order to incorporate the user meta-preferences M into the simulation, we had to adjust the predicted user preferences $\hat{U}$ of the content-based recommender system.

As previously explained in section 3.2.2, the meta-preference vector M contains values between -1 and 1 and correspond to a user's preferred direction of preference shift. The format of the meta-preference vector lends itself quite well to simply adding M to $\hat{U}$. We also thought it was important to introduce a hyperparameter α which determined how strongly the recommender system would incorporate M. Because the meta-preferences are incorporated after the model is trained on, this would be considered a post-processing method. To accomplish this in T-RECS, we implemented a custom content-based recommendation model which is the same as the one formalized by Chaney et al. [2018] except that it incorporates the meta-preferences at each timestep after the startup-and-train phase is complete ($t = 50$ in our simulation).

3.3.2 Measuring Predicted-Target Similarity

In order to evaluate how effective the incorporation of the meta-preferences was, we measured the average cosine similarity of the predicted user profiles $\hat{U}$ and the target user profiles U_{target} over the course of the simulation.

Cosine similarity is a basic measure of similarity between two vectors. It is symmetric, bounded between -1 (most dissimilar) and 1 (exactly the same), and is calculated as

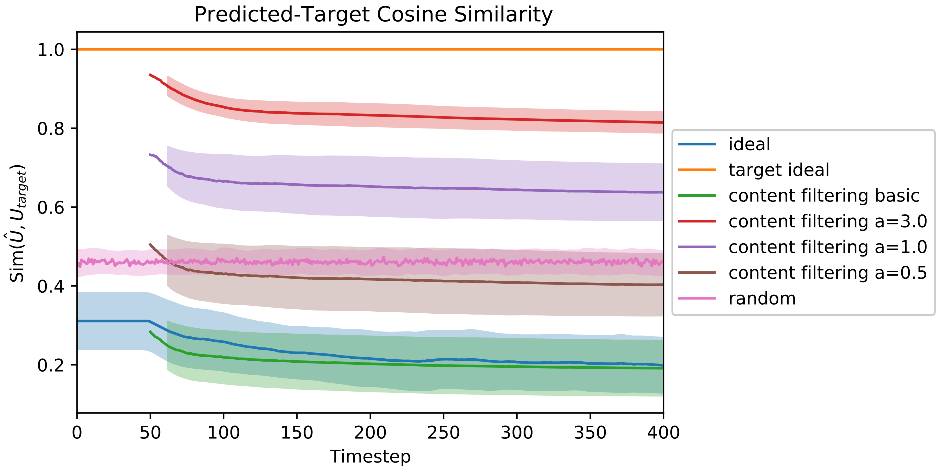
$$\text{Sim}(\hat{U}, U_{target}) = \frac{\hat{U} \cdot U_{target}}{\|\hat{U}\| \cdot \|U_{target}\|}.$$

To generate our results in T-RECS we implemented a custom metric which measured the similarity at each timestep in the simulation for all models. It was not possible to measure the similarity with the popularity model since the user representation $\hat{U}$ used by the popularity model was not attribute-based and therefore incomparable to U_{target}.

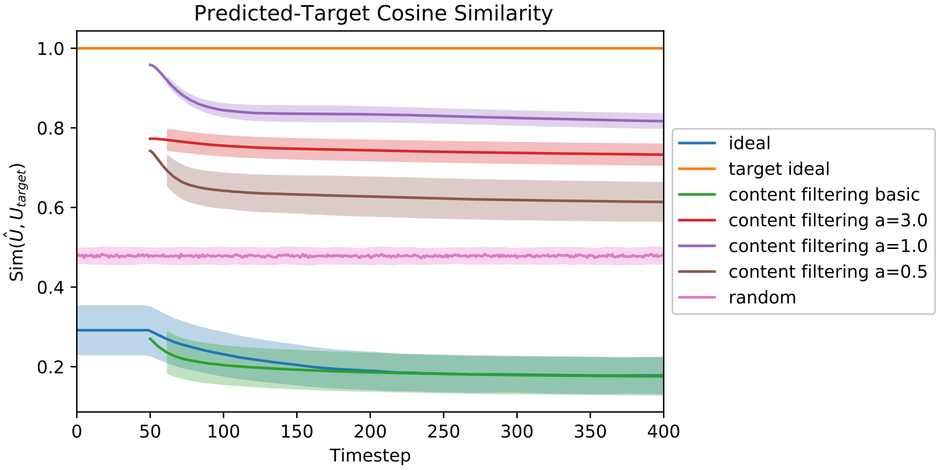
3.3.3 Results

Our results show that models which incorporate user meta-preferences have predicted user preferences which are much more similar to the user's target preferences than the random, ideal, and basic content filtering model.

Importantly, we noticed that within the content-based recommender system, the predicted preferences of each user were not normalized and therefore did not align with our conception of user preferences as a discrete probability distribution. After testing the simulation with and without normalized $\hat{U}$ in all the models, we decided to normalize the profiles because it improved the effectiveness of the meta-preferences, permitted α values larger than 1.0, and had no impact on the actual behavior of the recommender systems when presenting users with items. The graphs are presented in 3.7 and we see that



(a) Predicted-Target Similarity with non-normalized preferences. The model with strong incorporation ($\alpha = 3.0$) has the highest similarity, followed by exact ($\alpha = 1.0$). The model with weak incorporation ($\alpha = 0.5$) is less similar than the random model.



(b) Predicted-Target Similarity with normalized preferences. The most similar model is the exact ($\alpha = 1.0$) followed by strong ($\alpha = 3.0$) and weak ($\alpha = 0.5$). All extended content models are more similar than the random model.

Figure 3.7: Predicted-Target Similarity for normalized and non-normalized preferences. We see that the extended content models achieve higher similarity with normalized preferences, with the exception of the strong model with $\alpha = 3.0$ which is slightly less similar than with non-normalized profiles. The popularity model is not included because the representation of $\hat{U}$ was not comparable to those of the other models. Results are averaged over 25 independent simulations. A version of these plots with a log-log scale are included in the appendix in figure A.1.

the similarity is substantially higher for the simulation with the normalized profiles.

For the normalized preferences in figure 3.7b we see that, aside from the target ideal model which uses a user representation which is exactly the same as U_{target}, the extended model with the "exact" meta-preferences $\alpha = 1.0$ leads to the highest overall similarity throughout the simulation. This is then followed by the model with the strong meta-preferences $\alpha = 3.0$ and weak meta-preferences $\alpha = 0.5$. High values of α which push $\hat{U}$ to be more extreme in the direction of the meta preferences does lead to a lower similarity, but may result in better results in terms of user-directed preference change as discussed in section 3.5. The random recommender system lingers at a similarity of 0.5 as expected since its $\hat{U}$ is randomly generated at each timestep. The ideal recommender system uses the actual user preferences U as its predicted user preferences $\hat{U} = U$ and here we see that the actual user preferences slowly drift to become more dissimilar to U_{target} as they are served items which perfectly align with their actual preferences. Finally, the basic content-based recommender systems are the worst of all the models in regards to the predicted-target cosine similarity.

All of the models show a dip in similarity before stabilizing. This is due to the preference drift mechanism, which leads many of the smallest user attributes to drop to 0 very quickly, slightly homogenizing their preferences and lowering the similarity. After a certain point this homogenization stabilizes. This phenomenon is discussed in more detail in 3.4.

3.3.4 Discussion

From the results, we see that our method of incorporation was effective in aligning the models' predicted user preferences with the user's target preferences. Although the extended content models are more similar to the user's target preferences, their recommendations may be less relevant to the actual user preferences U, which are the basis for user-item interaction. Although in this simulation, users consume an item at every timestep, in a system predicated on engagement or satisficing choice, this could lead to a less accurate and engaging recommender system. The effects of the meta-preference mechanism on accuracy and engagement are discussed more in section 3.6.

The method we used to incorporate the meta-preferences into $\hat{U}$ only works for the content-based model. Other recommender systems in T-RECS do not represent users in a way which is compatible with a content-based meta-preference mechanism. Further research should be done on how other kinds of recommender systems could facilitate user-directed preference change.

Now that we have successfully incorporated M into the models, it is time to begin investigating what effect this will have on user preferences as the models recommend items aligned with these meta-preferences.

3.4 RQ3: Recommendations and Preference Change

Before looking at the effect of our mechanism on user-directed preference change, we must first investigate how to simulate preference change within the system. As discussed in section 2.2, preferences will adapt to environmental influences, including that of the recommended items they consume. Simulations which assume users have underlying “true” preferences which are not affected by recommendations miss an incredibly important aspect of recommender system dynamics and are in danger of creating systems which affect user preferences in unknown and undesired ways.

In this section we answer the following sub-research question:

> How can we simulate preference change among users?

We first present the basic mechanism we used to drift user preferences after each item interaction in section 3.4.1. We then discuss the homogenization of user preferences throughout the simulation and how we measured and reduced it by using a fixed item catalog in section 3.4.2. We present our results demonstrating preference change within the simulation for different models in section 3.4.3 and discuss these results in section 3.4.4.

3.4.1 Simulating Preference Drift in T-RECS

To simulate preference change in the simulation, we used the existing mechanism in T-RECS for dynamic user preferences with a drift parameter of $d = 0.02$. Lucherini et al. [2021] implemented this basic model of preference drift with spherical linear interpolation [Shoemake, 1985], where the U vector rotates in the direction of the attribute of the chosen item. In the simulation, after a user has chosen the item from their recommended set with the highest estimated utility (subject to the attention mechanism described in 3.1.1) their preferences drift in the direction of that item’s attribute. This is visualized in figure 3.8.

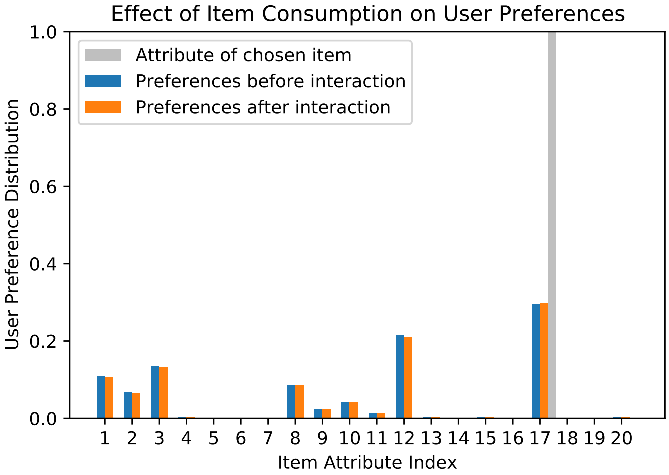


Figure 3.8: Actual user preferences U are drifted with drift parameter $d = 0.02$ toward the attribute of the chosen item.

Our choice of drift parameter $d = 0.02$ was informed by computational power and the length of our simulation. During our experiments, we found that using higher or lower values for d did not change our results significantly, but it did condense or expand the graphs along the x-axis. In our final simulation, we used $t = 50$ timesteps for the start-up phase and $t = 350$ timesteps for the recommendation phase. With our chosen drift parameter, this was enough to show stabilization in the Actual-Target Similarity as presented in section 3.5 as well as stabilization in the homogenization effects discussed in the next subsection 3.4.2. The user profiles do not drift in the start-up phase in order to keep the features of the preferences at the beginning of the recommendation phase consistent with the parameters by Chaney et al. [2018].

The limitations of this model of preference change are discussed in section 3.4.4.

3.4.2 Reducing User Preference Homogenization

One consequence of using the dynamics of a simulation used to study homogeneity was, unsurprisingly, strong homogenizing effects on user preferences. Because the main objective for this thesis is to investigate user-directed preference change, strong and immediate homogenization of user preferences made the effect of the meta-preference mechanism difficult to discern.

In figure 3.9, we see that, even for the random recommender, the preferences U of a randomly chosen user gradually homogenize to become more concentrated around only a few attributes when using the dynamic content-creator catalog of Lucherini et al. [2021].

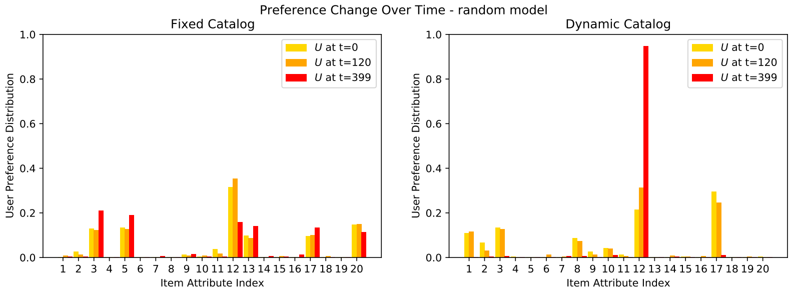


Figure 3.9: Preference change over time of a random user subject to random recommendations. The preferences with the fixed catalog are much less visibly homogenized by the end of the simulation ($t = 399$) than with the dynamic catalog. The preferences at $t = 120$ are shown to demonstrate the distribution after the Actual-Target Similarity has stabilized but before strong homogenization has set in.

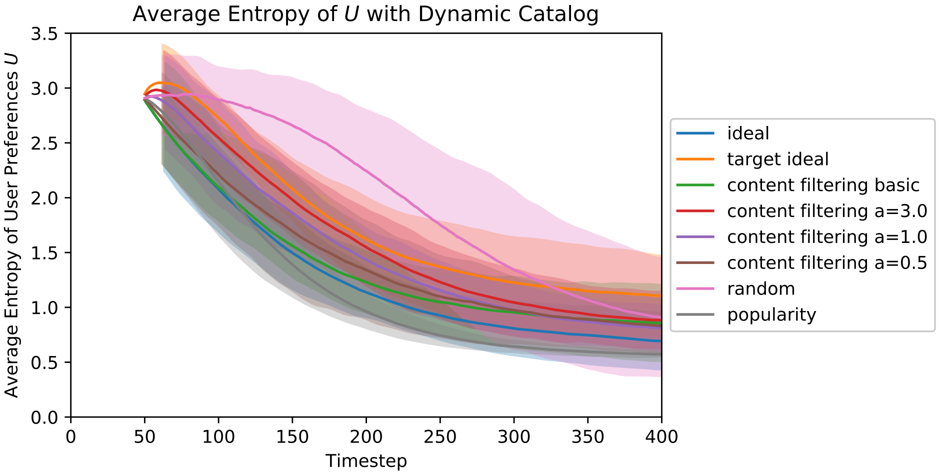
In order to see what effect the meta-preference mechanism had on user preferences, we had to relax the homogenization effects of the larger system. To do so, we explored using a catalog generated from a fixed distribution like that of Chaney et al. [2018], rather than the adaptive creator distributions of Lucherini et al. [2021]. The fixed catalog ensured that the items available to be recommended by the models were not increasingly homogenized, so any homogenizing effects on user profiles could be more concretely attributed to the models, rather than the catalog.

To measure the homogenization of individual user preferences and compare them for the fixed and dynamic catalogs, we used Shannon entropy as a measure of user preference homogenization. This choice of metric was inspired by the use of a similar metric to assess content creator homogenization in Lucherini et al. [2021]. Entropy is generally a measure of uncertainty, with higher entropy values corresponding to more "uncertain" probability distributions [Bromiley et al., 2004]. It is calculated in our simulation as

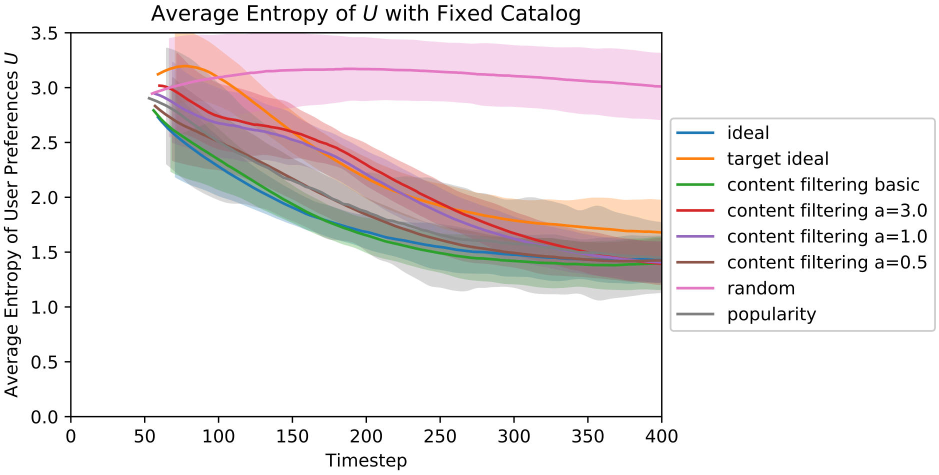
$$H(U) = -\sum_{i=1}^{A} U_i \log_2 U_i.$$

If the entropy of user preference distribution U is high, then the profiles are more balanced and the user is more likely to be choose items of many different attributes (we are less certain which item a user will tend to choose). If the entropy is low, then the user will tend to choose items with the same few attributes. To measure this in T-RECS we implemented a custom metric which averages Shannon entropy of all actual user preferences U at each timestep. The results of this are presented in Figure 3.10.

Comparing the graphs for the fixed and dynamic catalogs in figure 3.10, we can see that the simulation with the fixed catalog leads to less homoge-



(a) Average entropy of user profiles using the dynamic content creator catalog like that of Lucherini et al. [2021]. Results are averaged over 15 independent simulations.



(b) Average entropy of user profiles using the fixed content creator catalog like that of Chaney et al. [2018]. Results are averaged over 25 independent simulations.

Figure 3.10: Homogenization of user preferences with fixed vs dynamic item catalogs. User preferences experience less homogenization overall with an item catalog generated from a fixed distribution. Because these plots were for purposes of direct comparison, the log-log plots are not included in the appendix.

nization overall. This is also evident in the profile plots in figure 3.9, where the preferences at the end of the simulation with the random recommender were much less homogenized with the fixed catalog. For this reason, we chose to use the fixed catalog in our simulation. The consequences of this choice are discussed below in section 3.4.4.

Interestingly, the expanded content models lead to the least amount of individual profile homogenization. One explanation for the success of the models with the meta-preference mechanism comes from Jiang et al.'s [2019] observation that noise in the prediction accuracy of a recommender system will dampen homogenization effects on preferences. Whereas Jiang et al. [2019] tested this with random noise, the meta-preferences do act as a dampening agent to slow down the degeneracy of the homogenizing feedback loop in the system. The effects of the meta-preference mechanism on system-wide homogenization are discussed further in section 3.6.1.

3.4.3 Results

Using the same metric as Jiang et al. [2019] to measure the degree of preference change over time, we found that the models with the meta-preference mechanism lead to the strongest preference change.

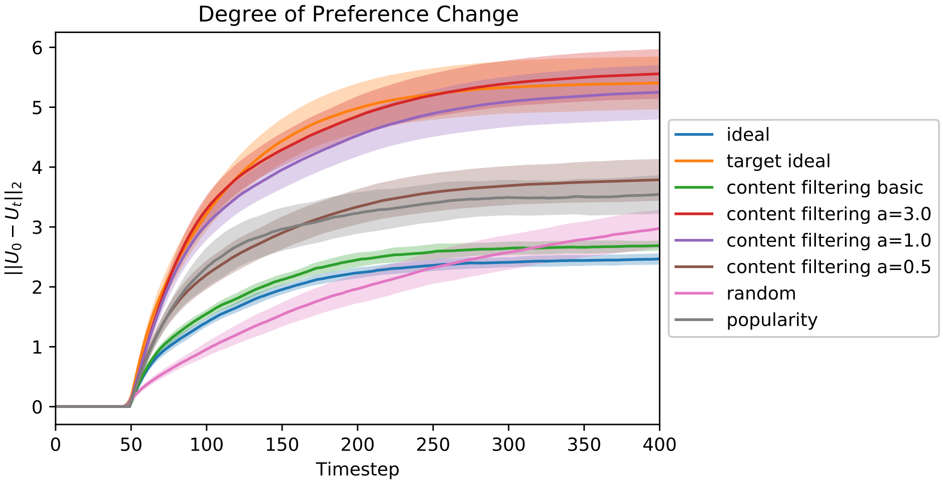


Figure 3.11: User preferences U in the simulation experience a fast process of preference change and eventual stagnation. The target ideal and content models with the strong and exact meta-preference mechanism experienced the highest degree of change. The random, ideal, and basic content models lead to the least degree of preference change. A version of this plot with a log-log scale is included in the appendix in figure A.2.

The L2 norm used by Jiang et al. [2019] is useful in measuring the speed and intensity of preference change, but the direction of the change is not visible. It is impossible to determine the degree of user-directed preference change from these results.

3.4.4 Discussion

Our model of preference change is simple and useful for simulating basic preference drift, but is far from complete or comprehensive. It only takes into account the items chosen by the user, whereas it is also possible that the whole recommended set may influence user preferences in some form. Additionally, it assumes equal impact of all items and attributes on user preferences. The research discussed in section 2.2.3 finds that some kinds of content, particularly addicting content and items which foster dependency, are more prone to cause preference change. A more sophisticated model of user preference change could have been implemented, such as the one proposed by Bernheim et al. [2021], but we decided to use the default model in T-RECS for purposes of simplicity.

Our use of a fixed item catalog rather than a dynamic one brings attention to the need to address the issue of preference homogenization in a system with item homogenization. The choice of catalog does not have large consequences for the ultimate findings of our research, but does imply that the kind of system we are simulating is one where the items are not adaptively catered to user behavior. It could also be interpreted that the timescale of our preference change simulation is not relevant to systemic changes in item attributes. Investigating the interaction of user-preference change and creator adaptation would be an interesting angle for future work.

Our measure is suitable for identifying the degree of preference change caused by different recommender models, but not the direction. Identifying the direction of preference change is integral to studying user-directed preference shift and distinguishing between wanted and unwanted shifts. The next section will look at the alignment of the changed user preferences with U_{target}.

3.5 RQ4: User-Directed Preference Change

Now that we have delineated all the relevant dynamics of the system, it is time to measure user-directed preference change within our simulation.

In cases where a user's behavior is at odds with another preference, most recommender systems are only capable of catering to behavior and therefore miss the opportunity to serve users in a deeper, more fulfilling way. Having operationalized this conflict in our simulation as a user's actual preferences U vs their target preferences U_{target}, we now look at the relationship as the simulation evolves under different models.

In this section we answer the following sub-research question:

> Do recommendation models which incorporate user meta-preferences lead to preference change aligned with their target preferences?

We begin by discussing our use of cosine similarity to measure the alignment of the actual and target user preference distributions in section 3.5.1. The results are presented in section 3.5.2 and demonstrate higher similarity for models which incorporate M. The direct implications and limitations of these results are discussed in 3.5.3.

3.5.1 Measuring Actual-Target Similarity

Instead of testing the similarity of the predicted user preferences $\hat{U}$ held by the model, we now measure the similarity of the actual user preferences U and the target user preferences U_{target} throughout the simulation. Similarly to section 3.3, we implemented a custom metric in T-RECS to measure the cosine similarity of the actual and target preferences, $\text{Sim}(U, U_{target})$ averaged over all users. The properties of cosine similarity are discussed in 3.3.2.

3.5.2 Results

Our results show that models which incorporate user meta-preferences M lead to preference change in users which is much more similar to their target preferences than the preference change caused by other models.

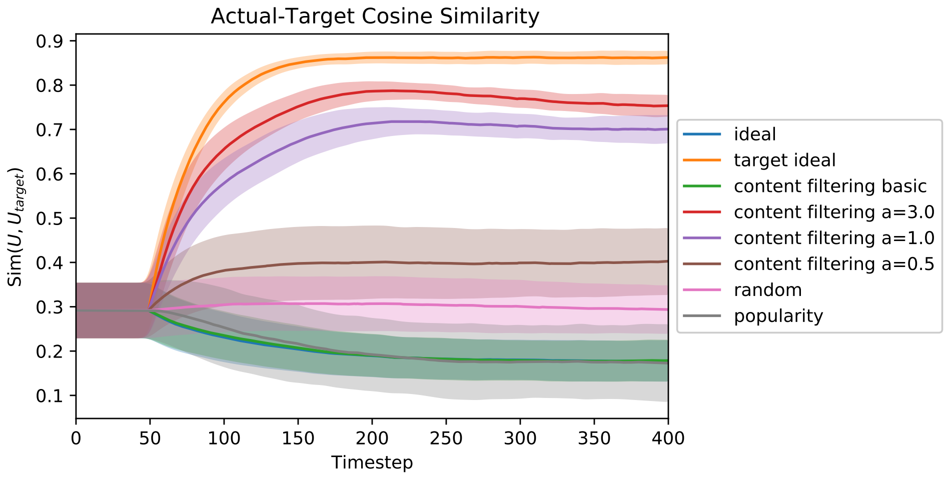


Figure 3.12: The average Actual-Target Cosine Similarity over time. Actual user preferences U do not drift in the start-up phase, but rapidly become more similar in the models which incorporate user meta-preferences M. The random recommender leads to little preference change. The ideal, basic content, and popularity models lead to preference change which is increasingly dissimilar to the user's target preferences U_{target}. Results are averaged over 25 independent simulations. A version of this plot with a log-log scale is included in the appendix in figure A.3.

The target ideal model serves recommendations which are perfectly aligned with user target preferences, which would not be available to actual recommender systems with partial knowledge. It is therefore the most effective for the purposes of user-directed preference change, but is unable to achieve complete similarity. This is due to the homogenization effects discussed in section 3.4.2, as well as the coarseness of the preference change model.

The most successful model is the content filtering recommender which strongly incorporates M with $\alpha = 3.0$. This may be surprising, since the exact model ($\alpha = 1.0$) gives a more accurate picture of the user's wishes and had a higher Predicted-Target Similarity, but over-incorporating M leads to models which provide top-N recommendations which are less catered to the user's actual preferences U, and therefore less likely to reinforce those preferences as U drifts toward the user's chosen items. The exact model with $\alpha = 1.0$ is also quite effective at facilitating user-directed preference change compared to all other models. The model with weak meta-preference incorporation with $\alpha = 0.5$ led to much slower and less pronounced user-directed preference change, but was still better than the random model. Although the strong and exact models have higher Actual-Target Similarities, there is a trade-off between more effective preference change and other metrics like accuracy and engagement. This is the focus of the following sub-research question in section 3.6.

The random recommender demonstrates preference change which does not change the similarity in any systemic way, which was expected. The visualization in figure 3.13b demonstrates this for a randomly selected user. The decreasing similarity of the ideal, basic content filtering, and popularity models indicate that those models cause user preferences to shift *away* from their target ideal at around the same rate. This demonstrates why it is so important for models to consider user meta-preferences and prevent preference change which is *not* sensitive to the user's wishes.

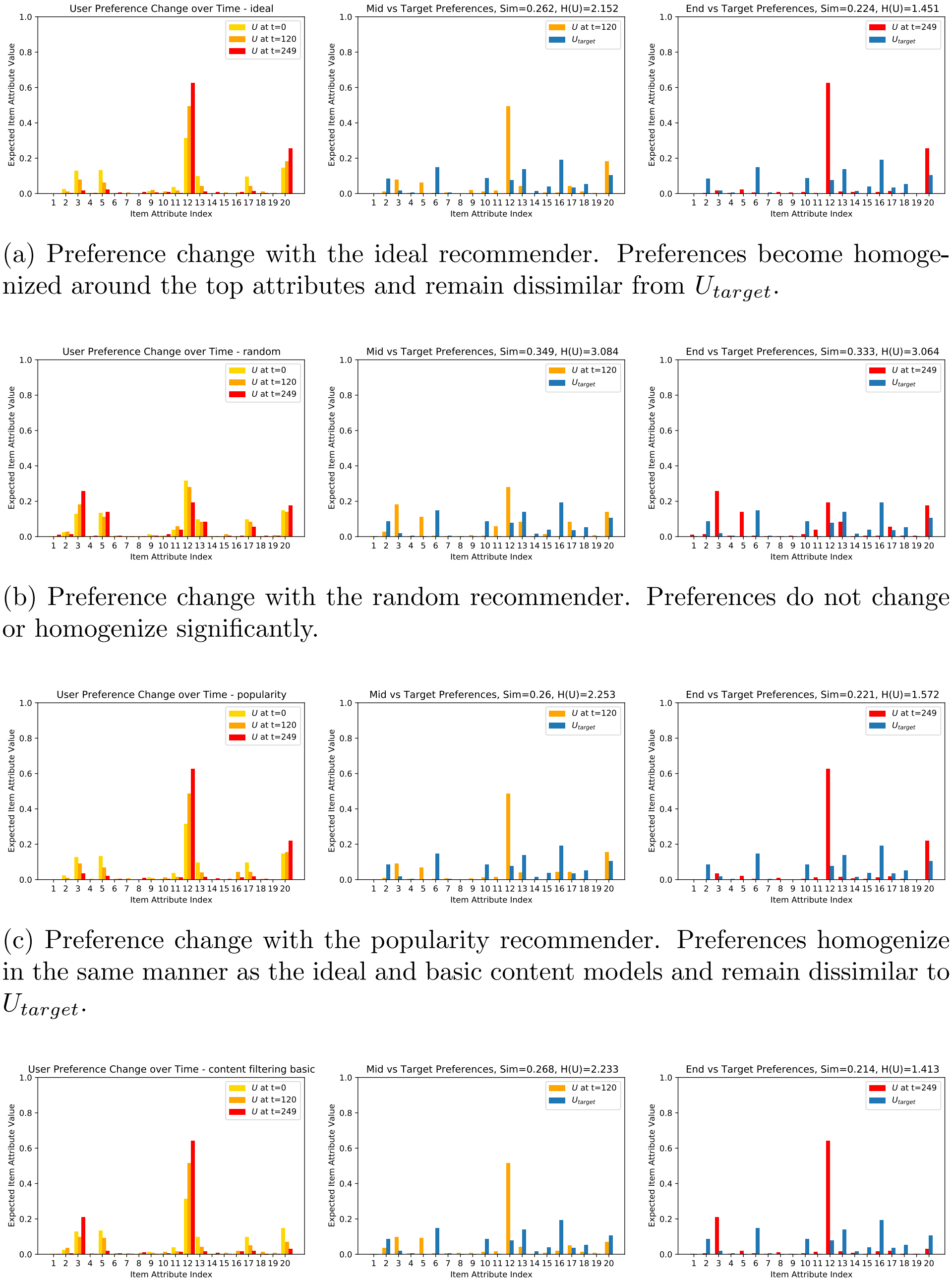
Figure 3.13 shows how a random;y selected user's preferences change over time subject to different recommendation models. Because the Actual-Target Similarity stabilizes before the end of the simulation we plot the user profiles in the middle (taken at $t = 120$) and at the end of the simulation to showcase the preferences at a point before the homogenizing effects set in. The plots correspond to the similarity findings and also portray the homogenizing effects of the recommender systems, causing strong preferences to become stronger and weak preferences to diminish over time to different degrees. The plots also demonstrate that the incorporation mechanism and similarity measures are not especially exact and cannot express the success of attributes individually.

3.5.3 Discussion

Our research shows promising results for the possibility of user-directed preference change in recommender systems and demonstrates the efficacy of various models in facilitating such change.

As demonstrated by the ideal, basic content-filtering, and popularity models, systems which do not incorporate user meta-preferences and preference change preference inevitably lead to preference change which could be misaligned with user target preferences. The implications of this are important to fixing the preference change blindness present in most present-day recommender systems and simulations. If a system is deployed circumstances relevant to preference change, the effects on users should at least be acknowledged and ideally addressed. For this, techniques for measuring preference change in recommender systems should be introduced and standardized. The work of Sanna Passino et al. [2021] in modeling dynamic user preferences in existing data could be useful, and the larger field of Preference Science [Franklin et al., 2022] will hopefully begin to answer these questions in the coming years.

Our results unexpectedly linked the issue of preference change to homogeneity in recommender systems. Because U_{target} was generated as a stationary distribution with relatively low homogeneity, the models which incorporated M led to user preferences which were less homogenized. Correspondingly, as U became more homogenized in the models without the

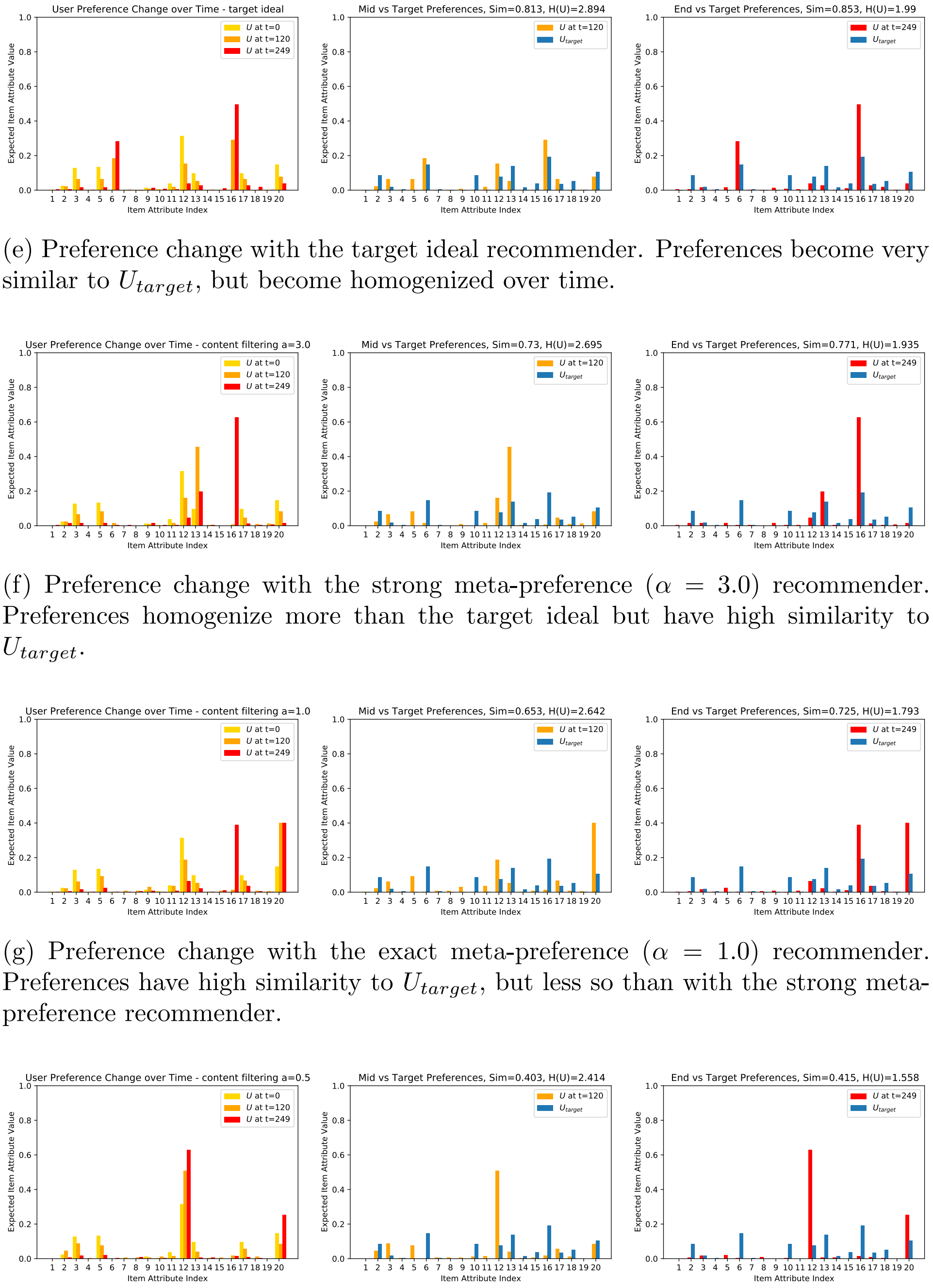


(a) Preference change with the ideal recommender. Preferences become homogenized around the top attributes and remain dissimilar from U_{target}.

(b) Preference change with the random recommender. Preferences do not change or homogenize significantly.

(c) Preference change with the popularity recommender. Preferences homogenize in the same manner as the ideal and basic content models and remain dissimilar to U_{target}.

(d) Preference change with the basic content filtering recommender. Preferences homogenize in the same manner as the ideal recommender and remain dissimilar to U_{target}.



(e) Preference change with the target ideal recommender. Preferences become very similar to U_{target}, but become homogenized over time.

(f) Preference change with the strong meta-preference ($\alpha = 3.0$) recommender. Preferences homogenize more than the target ideal but have high similarity to U_{target}.

(g) Preference change with the exact meta-preference ($\alpha = 1.0$) recommender. Preferences have high similarity to U_{target}, but less so than with the strong meta-preference recommender.

(h) Preference change with the weak meta-preference ($\alpha = 0.5$) recommender. Preferences have higher similarity with U_{target} than models without meta-preference incorporation but are more homogenized and less similar than the strong and weak models.

Figure 3.13: Examples of user preference change trajectories over time subject to various recommendation models. Preferences at the middle and end of the simulation are plotted with corresponding similarity and entropy scores to visualize stable similarity and stable homogeneity effects respectively.

meta-preference mechanism, the Actual-Target Similarity decreased since U_{target} could only have limited cosine similarity with highly homogenized U. For this reason, a plausible representation of U_{target} is very important to investigate in future research. It would also be valuable to investigate meta-preferences in a system controlled for homogeneity, possibly through the use of weighting techniques like Schnabel et al. [2016] and De Myttenaere et al. [2014]. This would give a clearer idea of the direction and nature of preference change in models which do not incorporate M.

We have now measured the direct effect of meta-preference incorporation on user preferences and briefly explored the homogenization of those preferences (section 3.4), but what effects might M have on the characteristics of the recommendations given to users or on the system as a whole? We will explore these important questions in the final research section of this thesis.

3.6 RQ5: Further Effects of User-Directed Preference Change

A central benefit of simulation is the ability to study the dynamics of the simulation with a variety of other metrics. Although we were primarily interested in investigating the effect of the meta-preference mechanism on user-directed preference change, it is probable that such a mechanism might lead to various other second-order effects on recommendations which need to be anticipated. Though our simulation is far from having strong ecological validity, we can still demonstrate the relationship of such a mechanism in a broad sense, and use these insights to outline further areas of future research.

In this section, we answer the following sub-research question:

> What are the some effects of user-directed preference change on recommendations?

We investigate the effect on recommendations by applying some popular metrics for accuracy, diversity, between-user homogenization, and hypothesize the possible effects on engagement in section 3.6.1. We then discuss potential societal impacts and risks of a meta-preference mechanism in section 3.6.2, and discuss some limitations of our solution as a solely individual bottom-up approach.

3.6.1 Effects on Recommendations

Accuracy

There exists a trade-off in accuracy and the strength of meta-preference incorporation.

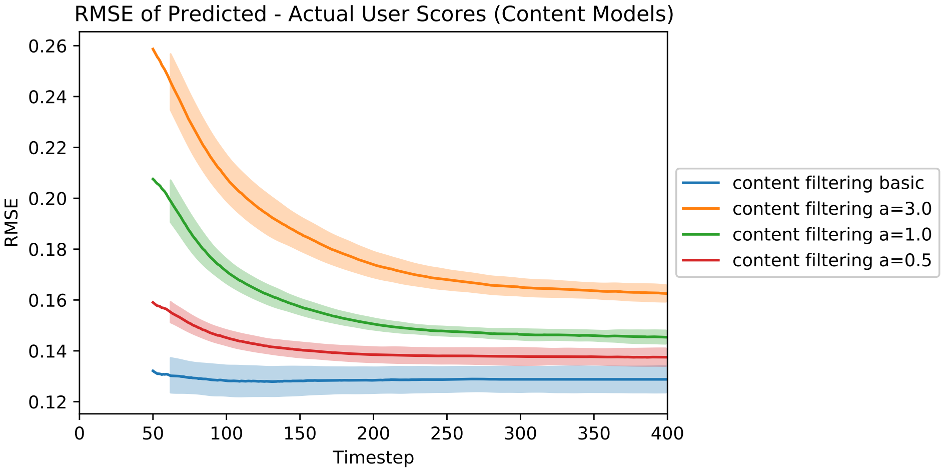


Figure 3.14: RMSE of the content filtering models increases as the strength of meta-preference incorporation rises. Results are averaged over 25 independent simulations. A version of this plot with a log-log scale is included in the appendix in figure A.4.

In T-RECS the root mean squared error (RMSE) metric is measured using the mean difference between the predicted user scores $\hat{S}$ (taken as the dot product between the item attributes I and the predicted user preferences $\hat{U}$) and actual user scores S. This allows T-RECS to report an accuracy metric despite simulating recommender systems trained on implicit feedback. It is calculated as

$$RMSE(S, \hat{S}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{S}_i - S_i)^2} \tag{3.1}$$

where N is the number of users.

In figure 3.14 we see the RMSE scores for the content models. As the strength of the incorporation parameter α increases, we see that the RMSE does as well. This would indicate that more successful user-directed preference shift comes with a trade-off in accuracy.

The random and popularity models are not shown in figure 3.14 because their scores were so much higher than the content models. The RMSE of the random model stabilized at a value of 2.0 and the popularity model stabilized around a value of 26. The ideal models are also not included since they do not serve recommendations based on user scores.

Diversity

Using the the metric of interaction spread as a measure of diversity, we found no significant difference in the diversity of any of the content models.

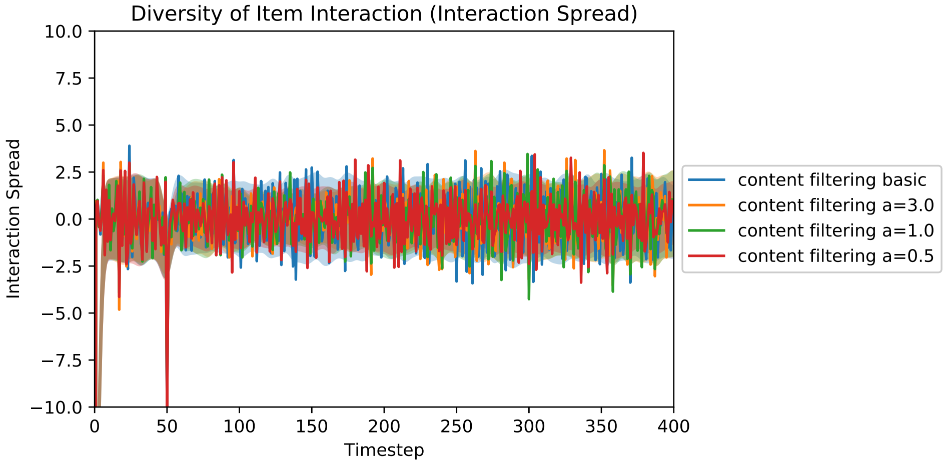


Figure 3.15: No systemic differences in interaction spread are observed for any of the basic vs. extended content models. The structure of this graph did not translate to an interpretable log-log plot, so it is not included in the appendix.

The interaction spread metric is a base metric in T-RECS and measures whether interactions are spread among many or just a few items. We do note that this metric is fairly difficult to read from graphs, but 3.15 does not demonstrate a large divergence between models.

Homogenization

When investigating the effect of the meta-preferences on between-user homogenization, we found that the models with the strong meta-preferences led to less between-user homogenization than all other models.

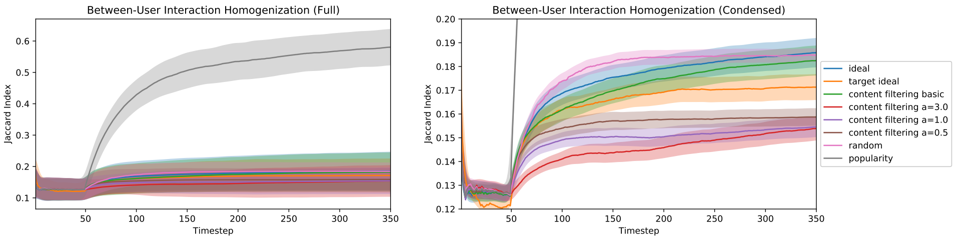


Figure 3.16: The average Jaccard similarity between interacted items at each timestep as a measure for between-user interaction homogenization. We see the between-user homogenization very high for the popularity model as expected, and is less for the models which incorporate user meta-preferences. Results are averaged over 25 independent simulations. The standard deviation of the error bars has been reduced from 1.0 to 0.1 for interpretability in the condensed plot. A version of this plot with a log-log scale is included in the appendix in figure A.5

We use Chaney et al.'s [2018] metric of homogenization which measures the average Jaccard similarity between the attributes of the chosen items at each timestep. Our results in figure 3.16 differ from the findings of those of Chaney et al. and Lucherini et al. [2021] in a few interesting ways, notably that the random and ideal models have of the highest degrees of homogenization with the exception of the popularity model, which is much more homogenized than in the original studies. These differences could potentially be explained by the presence of preference drift in this simulation, which Chaney et al.'s simulation did not include. Further research into the effect of preference shift mechanisms on between-user homogenization would be valuable.

Our results correspond with our observations about within-user preference homogenization, and the observation that noise or inaccuracy within the predictive model of a recommender system leads to less homogenization overall [Jiang et al., 2019]. The same limitations discussed in section 3.4.4 also apply. The lower homogeneity probably occurs due to the meta-preference mechanism's ability to increase the similarity of U to the less-homogenized target distribution U_{target}. If the target preferences were not as uniform, it is unlikely we would observe these effects, so the validity of how U_{target} is simulate should be researched further.

Engagement

Our simulation did not study the effect of meta-preferences on engagement since users interacted with items every timestep as per the dynamics of our simulation. However, we still believe that engagement is a relevant system-wide effect to mention in this section, since it is one of the dominant paradigms of commercial recommender system evaluation.

We take engagement to mean any combination of click-rate, retention, time-on-site, and other so-called “captivation metrics” [Seaver, 2019] which are used to optimize user behavior to meet the directives of the recommender system designers. With the trade-off in user-directed preference change and accuracy, it is not unreasonable to believe that any model which facilitates user-directed preference change may lead to a recommender system which is less engaging, particularly with entertainment-based recommender systems. While this may lead to more enriching experience for the user and even allow for better control over time spent using the system, it may be problematic for commercial systems with an engagement directive tied to earnings.

As we discussed in 2.3.3, though the theoretical grounding and utility of a mechanism like this may be solid, the actual applications are limited until the values of technologies like recommender systems can be aligned to encompass broader goals of human and societal development.

3.6.2 Societal Impacts and Risks

Though we have taken user-directed preference shift to be a positive use of recommender systems, there exist significant risks and misuses of the technology which need to be anticipated and prevented in a real-world implementation.

If the parameters of the target user preferences U_{target} were set by someone other than the user, then the recommendation system could be used as a tool for subversive preference manipulation rather than one for self-development. Providing an easy mechanism for content-based curation could potentially make this kind of attack easier.

Additionally, the ability to direct their own preferences could provide people who are already within echo chambers with the means to isolate themself from alternative viewpoints and resist any attempts at corrective diversity or depolarization by the recommender system [Stray, 2021].

Our simulation is also limited in its conception of user-directed preference change as an individual-level phenomenon. Conceiving of self-development as an insulated and non-social process ignores the benefits and importance of societal cohesion and coordination. In this case, leaving it up to the individual may be an easy way to sidestep the ethical issues of centralized top-down planning, but ignoring the possibility for collective coordination and group socio-cultural development could be problematic.

The power of systems which can cause preference change cannot be understated. Systems blind to it are dangerous in their inability to know how they influence user preferences and what can be done about it. But it is also necessary that any approach to the problem of induced preference change must not be reactive or blind to second or third-order effects. The whole socio-technical environment must be considered to the best degree possible.

Chapter 4

Conclusion

This project was undertaken to introduce a solution to the problem of recommender system-induced preference change and evaluate the first and second-order effects of such a solution. Our research question asked,

> How can simulations of recommender systems incorporate users' meta-preferences to result in user-directed preference change?

To answer this question, we conceptualized meta-preferences as expressing the difference between a user's actual preferences and some distribution of target preferences. This formalization has several limitations, including the limited conceptualization of preferences as exclusively attribute-based as well as an unrealistic stability in target and the corresponding meta-preferences. We call for further research into the best way to simulate and realistically evoke meta-preferences from users.

We incorporated the meta-preferences by adding them to the model's representation of predicted user preferences at different strengths. We then implemented a custom metric to measure the similarity of the model's predicted user preferences and the user's target preferences to evaluate the effectiveness of our incorporation mechanism. We found that our method of incorporation worked best with normalized preferences and led to more similar predicted user preferences. This simple method of incorporation works within the simplified T-RECS models, but could not be used in real-world algorithms in a comparable way. Future research would be needed to find a compatible method of incorporation for real systems.

In simulating preference change with the basic model of preference drift in T-RECS, we found that extended content models led to the highest degree of preference change. We also observed that dynamic item catalogs led to much stronger within-user preference homogenization than fixed item catalogs regardless of model and opted to use fixed catalog to better isolate the effects of the meta-preference mechanism.

To measure user-directed preference shift in our expanded simulation, we used a custom metric to measure the similarity of the user's actual preferences and target preferences over time. We found that the model which has the strongest influence over the predicted user preferences is the most successful at resulting in user-directed preference shift, but comes with a trade-off in accuracy. We were not able to completely relax the homogenization effects but noticed that models with the meta-preference mechanism led to less homogenization due to the alignment with the user's minimally homogenized target preferences.

In studying further effects of the meta-preference mechanism, we demonstrated a trade-off between user-directed preference change and accuracy, found no effect on diversity, and observed less between-user homogenization due to target preference alignment. Despite these promising results, we acknowledge the challenges involved in the actual implementation of a mechanism like this, but strongly assert the need for recommender systems to address the preference change problem in a way that promotes user agency.

It is not enough to simply set up the infrastructure for user directed preference shift. To truly be a positive, value-aligned technology, the recommender systems should influence people only to be more self-directing, more self-aware, and to increase their capacity to make sense of the world and make good decisions.

Bibliography

Cecilie Schou Andreassen. Online social network site addiction: A comprehensive review. *Current Addiction Reports*, 2(2):175–184, 2015.

Guy Aridor, Duarte Goncalves, and Shan Sikdar. Deconstructing the filter bubble: User decision-making and recommender systems. *RecSys 2020 - 14th ACM Conference on Recommender Systems*, pages 82–91, 2020. doi: 10.1145/3383313.3412246.

Hal Ashton and Matija Franklin. The problem of behaviour and preference manipulation in ai systems. In *The AAAI-22 Workshop on Artificial Intelligence Safety (SafeAI 2022)*, 2022.

B. Douglas Bernheim, Luca Braghieri, Alejandro Martínez-Marquina, and David Zuckerman. A theory of chosen preferences. *American Economic Review*, 111:720–754, 2 2021. ISSN 19447981. doi: 10.1257/AER.20190390.

Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. Tasteweights: A visual interactive hybrid recommender system. pages 35–42, 2012. ISBN 9781450312707. doi: 10.1145/2365952.2365964.

PA Bromiley, NA Thacker, and E Bouhova-Thacker. Shannon entropy, renyi entropy, and information. *Statistics and Information Series (2004-004)*, 9, 2004.

Micah Carroll, Dylan Hadfield-Menell, Stuart Russell, and Anca Dragan. Estimating and penalizing preference shift in recommender systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 661–667, 2021.

Abhijnan Chakraborty, Gourab K. Patro, Niloy Ganguly, Krishna P. Gummadi, and Patrick Loiseau. Equality of voice: Towards fair representation in crowdsourced top-k recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 129–138. Association for Computing Machinery, 2019. ISBN 9781450361255. doi: 10.1145/3287560.3287570.

Allison JB Chaney. Recommendation system simulations: A discussion of two key challenges. *arXiv preprint arXiv:2109.02475*, 2021.

Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 224–232, 2018.

Arnaud De Myttenaere, Bénédicte Le Grand, Boris Golden, and Fabrice Rossi. Reducing offline evaluation bias in recommendation systems. *arXiv preprint arXiv:1407.0822*, 2014.

Katja De Vries. Identity, profiling algorithms and a world of ambient intelligence. *Ethics and information technology*, 12(1):71–85, 2010.

John Qi Dong. Using simulation in information systems research. *Journal of the Association for Information Systems*, 23:408–417, 1 2022. ISSN 1536-9323. doi: 10.17705/1jais.00743.

Daria Dzyabura and John R. Hauser. Recommending products when consumers learn their preference weights. *Marketing Science*, 38:417–441, 2019. ISSN 1526548X. doi: 10.1287/MKSC.2018.1144.

Michael D. Ekstrand and Martijn C. Willemsen. Behaviorism is not enough: Better recommendations through listening to users. In *Proceedings of the 10th ACM Conference on Recommender Systems*, page 221–224, 2016.

Michael D Ekstrand, Allison Chaney, Pablo Castells, Robin Burke, David Rohde, and Manel Slokom. Simurec: Workshop on synthetic data and simulation methods for recommender systems research; simurec: Workshop on synthetic data and simulation methods for recommender systems research. 3, 2021. doi: 10.1145/3460231.3470938.

Jon Elster. *Sour grapes*. Cambridge university press, 2016.

Matija Franklin, Hal Ashton, Rebecca Gorman, and Stuart Armstrong. Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of ai. 2022.

Batya Friedman, Peter H Kahn, Alan Borning, and Alina Huldtgren. Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory*, pages 55–95. Springer, 2013.

Daniel Geschke, Jan Lorenz, and Peter Holtz. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, 58:129–149, 2019. ISSN 20448309. doi: 10.1111/bjso.12286.

Lars Hall, Petter Johansson, and Thomas Strandberg. Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PloS ONE*, 7(9), 2012.

Md Rajibul Hasan, Ashish Kumar Jha, and Yi Liu. Excessive use of online video streaming services: Impact of recommender system use, psychological factors, and motives. *Computers in Human Behavior*, 80:220–228, 2018. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2017.11.020.

Anthony Jameson, Martijn C. Willemsen, Alexander Felfernig, Marco De Gemmis, Pasquale Lops, Giovanni Semeraro, and Li Chen. Human decision making and recommender systems. *Recommender Systems Handbook, Second Edition*, pages 611–648, 2015. doi: 10.1007/978-1-4899-7637-6_18.

Anthony Jameson, Martijn C. Willemsen, and Alexander Felfernig. *Individual and Group Decision Making and Recommender Systems.* Springer US, 2022. doi: 10.1007/978-1-0716-2197-4_21.

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate feedback loops in recommender systems. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Bart P. Knijnenburg, Saadhika Sivakumar, and Daricia Wilkinson. Recommender systems for self-actualization. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 11–14, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340359. doi: 10.1145/2959100.2959189. URL `https://doi.org/10.1145/2959100.2959189`.

Leonhard K. Lades and Liam Delaney. Nudge forgood. *Behavioural Public Policy*, 6:75–94, 1 2022. ISSN 2398-063X. doi: 10.1017/BPP.2019.53.

Yu Liang. Recommender system for developing new preferences and goals. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 611–615, 2019.

Eli Lucherini, Matthew Sun, Amy Winecoff, and Arvind Narayanan. T-recs: A simulation tool to study the societal impact of recommender systems. *arXiv preprint arXiv:2107.08959*, 2021.

Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Recommender systems and their ethical challenges. *AI and Society*, 35:957–967, 12 2020. ISSN 14355655. doi: 10.1007/s00146-020-00950-y.

Brent Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, 1(11):501–507, 2019.

Nicola Perra and Luis E.C. Rocha. Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific Reports*, 9, 2019. ISSN 20452322. doi: 10.1038/s41598-019-43830-2.

Steve Rathje, Jay J Van Bavel, and Sander van der Linden. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26), 2021.

Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Techniques, applications, and challenges. In *Recommender Systems Handbook*, pages 1–35. Springer, 2022.

Francesco Sanna Passino, Lucas Maystre, Dmitrii Moor, Ashton Anderson, and Mounia Lalmas. Where to next? a dynamic model of user preferences. In *Proceedings of the Web Conference 2021*, WWW '21, page 3210–3220, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450028. URL `https://doi.org/10.1145/3442381.3450028`.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR, 2016.

Julianne Schultz. Move very fast and break many things: Digital gangsters and the big other. *Griffith REVIEW*, (64):11–28, 2019. ISSN 1839-2954.

Nick Seaver. Captivating algorithms: Recommender systems as traps. *Journal of Material Culture*, 24:421–436, 12 2019. ISSN 14603586. doi: 10.1177/1359183518820366.

Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985.

Jonathan Stray. Designing recommender systems to depolarize. *CoRR*, abs/2107.04953, 2021.

Jonathan Stray, Ivan Vendrov, Jeremy Nixon, Steven Adler, and Dylan Hadfield-Menell. What are you optimizing for? aligning recommender systems with human values. *arXiv preprint arXiv:2107.10939*, 2021.

Rachel L Thomas and David Uminsky. Reliance on metrics is a fundamental challenge for ai. *Patterns*, 3(5):100476, 2022.

Sander Van der Linden. The conspiracy-effect: Exposure to conspiracy theories (about global warming) decreases pro-social behavior and science acceptance. *Personality and Individual Differences*, 87:171–173, 2015.

Wenjie Wang, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. User-controllable recommendation against filter bubbles. *ArXiv*, abs/2204.13844, 2022.

Amy A Winecoff, Matthew Sun, Eli Lucherini, and Arvind Narayanan. Simulation as experiment: An empirical critique of simulation research on recommender systems. *arXiv preprint arXiv:2107.14333*, 2021.

Appendix A

Appendix

In this appendix we include our results plotted with a log-log scale.

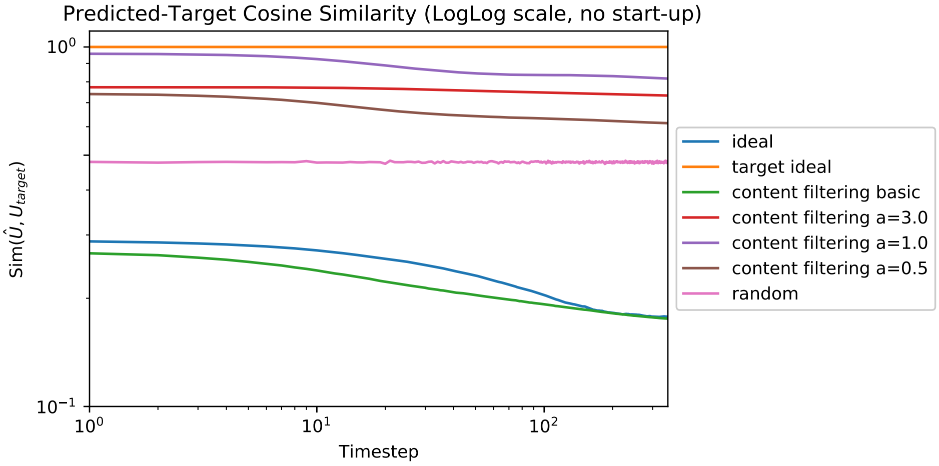


Figure A.1: Predicted-Target Similarity with non-normalized preferences. The model with strong incorporation ($\alpha = 3.0$) has the highest similarity, followed by exact ($\alpha = 1.0$). The model with weak incorporation ($\alpha = 0.5$) is less similar than the random model. Results are averaged over 25 independent simulations.

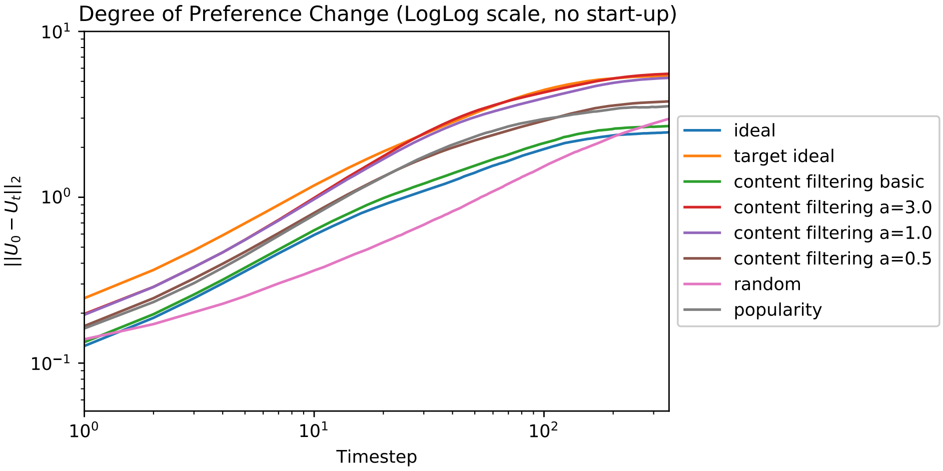


Figure A.2: User preferences U in the simulation experience a fast process of preference change and eventual stagnation. The target ideal and content models with the strong and exact meta-preference mechanism experienced the highest degree of change. The random, ideal, and basic content models lead to the least degree of preference change. Results are averaged over 25 independent simulations.

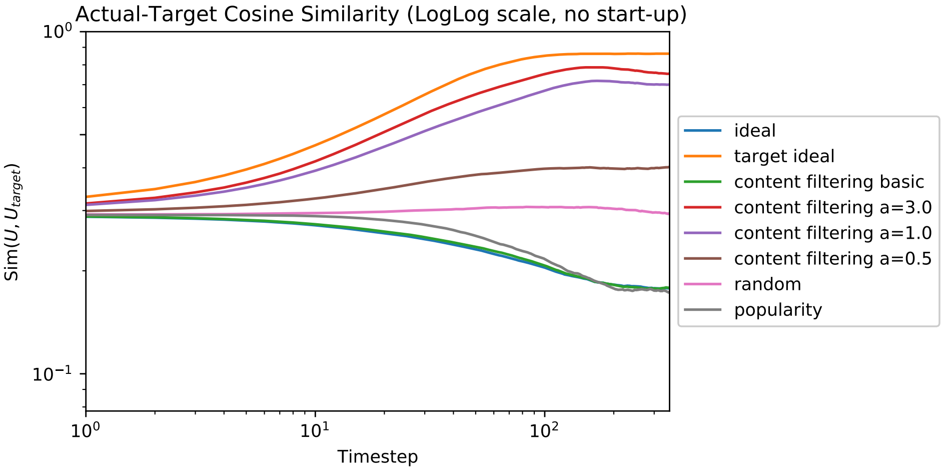


Figure A.3: The average Actual-Target Cosine Similarity over time. Actual user preferences U become more similar in the models which incorporate user meta-preferences M. The random recommender leads to little preference change. The ideal, basic content, and popularity models lead to preference change which is increasingly dissimilar to the user's target preferences U_{target}. Results are averaged over 25 independent simulations.

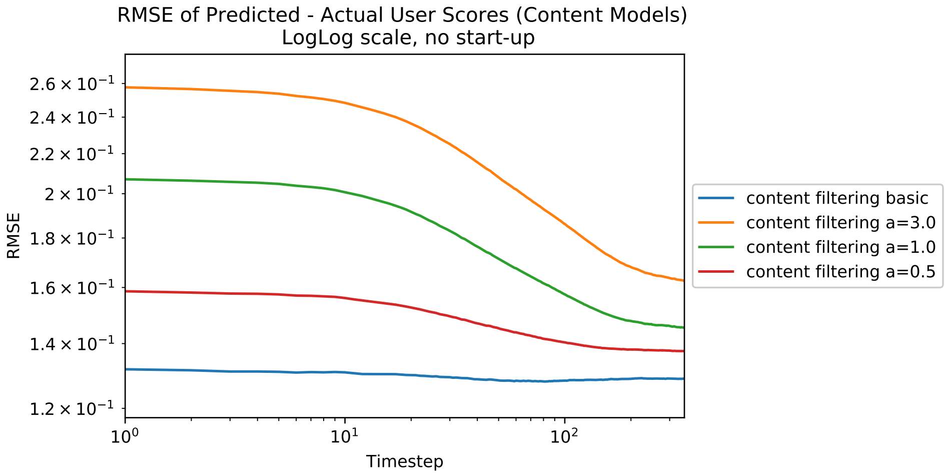


Figure A.4: RMSE of the content filtering models increases as the strength of meta-preference incorporation rises. Results are averaged over 25 independent simulations.

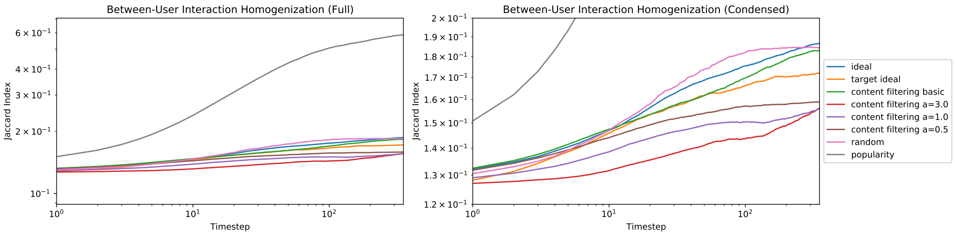


Figure A.5: The average Jaccard similarity between interacted items at each timestep as a measure for between-user interaction homogenization. We see the between-user homogenization very high for the popularity model as expected, and is less for the models which incorporate user meta-preferences. Results are averaged over 25 independent simulations. The standard deviation of the error bars has been reduced from 1.0 to 0.1 for interpretability in the condensed plot.