Amplification Pipelines

The Role of Feedback Loops in Recommender System Bias

Caio Lente

Thesis presented to the
Institute of Mathematics and Statistics
of the University of São Paulo
in partial fulfillment
of the requirements
for the degree of
Master of Science

Program: Computer Science
Advisor: Prof. Dr. Roberto Hirata Jr.

São Paulo
February 10th, 2021

Amplification Pipelines

The Role of Feedback Loops in Recommender System Bias

Caio Lente

This is the original version of the thesis prepared by the candidate Caio Lente, as submitted to the Examining Committee.

To my four parents and one and a half siblings.

Resumo

Caio Lente. **Viés Inescapável: *O Papel de Sistemas de Recomendação na Radicalização das Mídias Sociais***. Dissertação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2021.

Algoritmos de recomendação tornaram-se essenciais para o funcionamento de diversos sistemas que usamos no dia a dia, desde quais filmes assistir até quais produtos comprar. Entretanto, com a proliferação destes modelos nas redes sociais, surgiram também novas preocupações. Evidências anedóticas e um corpo cada vez mais robusto de pesquisa têm indicado que os algoritmos das redes sociais, por valorizarem engajamento, podem estar radicalizando usuários através da criação das chamadas câmaras de eco. Este trabalho pretende estudar algoritmos de recomendação como sistemas dinâmicos de modo a identificar se seus "espaços fásicos" estão sujeitos a dinâmicas de confinamento.

Palavras-chave: sistemas de recomendação. viés algorítmico. aprendizagem de máquina.

Abstract

Caio Lente. **Amplification Pipelines: *The Role of Feedback Loops in Recommender System Bias***. Thesis (Masters). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2021.

Recommendation algorithms have become essential to various day to day systems we use, from what movies to watch to what products to buy. However, with the proliferation of these models on social networks, new concerns have come to light. Anecdotal evidence and an ever growing body of research indicate that social network algorithms that promote engaging content might be radicalizing users, creating what has become known as echo chambers. The present study aims to study recommendation algorithms as dynamical systems as a means to identify if their "phase spaces" are subject to confinement dynamics.

Keywords: recommender system. algorithmic bias. machine learning.

Contents

1 Introduction **1**
1.1 Social Networks . . . 2
1.2 Recommender Systems . . . 2
1.3 Radicalization and Bias . . . 4
1.4 Research Goals . . . 5
1.5 Outline . . . 5

2 Literature review **7**
2.1 Scientific literature . . . 7
2.2 Journalistic efforts . . . 11

3 Static Analysis **13**
3.1 Datasets . . . 14
3.2 Experiments . . . 14

4 Dynamic Analysis **21**
4.1 Datasets . . . 22
4.2 Experiments . . . 22

References **27**

Chapter 1

Introduction

Social networks have all but taken over contemporary daily life. From the eponymous socializing, to reading news, to expressing ourselves, social media has creeped into every corner of society. Most of its side-effects, it could be argued, are positive (shortening distances, political accountability, social organizing), but they are not perfect institutions.

Social media companies already face significant backlash for their questionable business model and ethics. Cambridge Analytica's election meddling (), Facebook's subliminal experiments (), YouTube's problem with disturbing content marketed at kids (), and Twitter's bot infestation () are just a few recent scandals that have put the societal role of social media into question.

One particular controversy that has taken over public discourse around social networks is the role that their algorithms might have in radicalizing users, specially younger ones. The aforementioned experiments conducted by Facebook to influence people's emotions and the proliferation of more than questionable videos aimed at children on YouTube are instances that seem to corroborate the notion that there is something fundamentally wrong with these companies' algorithms.

News organizations, in general, have been skeptical of social networks. Journalists and specialists alike argue that social media's algorithms (specially recommender algorithms) are tuned to peddle conspiracy theories, extremist views, and false information (). This would be the source cause for a plethora of what they consider contemporary evils: religious extremism, anti-democratic leaders, widespread depression among teenagers, anti-science movements, etc.

This narrative, of course, has been questioned for a variety of reasons. Some say that it is self serving: traditional news organizations are being displaced by social media and it would be convenient for them to mine the public's trust in them (). Others claim that these recommender algorithms are not to blame for political polarization and that social networks even have a tendency to favor more left-wing viewpoints ().

The debate around the role of recommender systems in social media radicalization is still, unfortunately, too recent and based in anecdotes (). Since its impacts are all but

universal, more quality research is vital to inform both the public and opinion makers about if and how much recommendation algorithms influence social media users.

This dissertation aims to further such research.

1.1 Social Networks

Social networking services, also referred to as social networks and social media, are notoriously difficult to define. Some definitions might be too narrow (excluding instant messaging services), while some might be too broad (including technologies such as telephone networks). Most definitions () include some common features:

- Internet-based
- Focus on user-generated content
- Users have profiles
- Users can connect

While social-networking-like applications already existed in Usenet, Geocities, launched in 1994, is usually regarded as the first major social network. Friendster and Myspace followed in 2003, with Orkut and Facebook slightly lagging behind in 2004. Each hit their peak at different moments and different countries, but Facebook overtook all of them in 2009 when it became the most popular social networking service in the world, still maintaining the title over 13 years latter at the moment of writing ().

Even though all aforementioned social networks are multimedia, that is, users can post text, photos and videos, some of the most popular services focus on a specific type of media. For instance, YouTube (2009) centers around videos, WhatsApp (2009) and WeChat (2011) were originally designed for text-based communication, and Instagram's (2010) main focus is photos.

Parallel to all other features and idiosyncrasies, there lay the recommendation algorithms. While a few social networking services (e.g. WhatsApp) do not recommend any content or profiles to the user, most do and, according to recent studies, these recommendations have become the main drivers of interactions ().

1.2 Recommender Systems

Recommender systems (sometimes called recommendation systems or recommender algorithms) first appeared in 1992 under the name "collaborative filtering", even though that term nowadays refers to a subclass of recommender systems (Goldberg *et al.*, 1992). The aim of such an algorithm is providing users with personalized product or service recommendations, an essential task when considering the ever increasing number of possible videos to watch, music to listen, products to buy.

The input of a recommender system is usually information about the preferences (ratings, likes/dislikes, watch time, etc.) of consumers for a set of items. Preference information

can be gathered from explicit behaviors (e.g. rating a product in a scale ranging from 0 to 5 stars) or from implicit behaviors (e.g. how much time the user lingers on a product's page). These data can be combined with information about the user (age, political leaning, etc.) in order to create the best possible representation of the user's preferences ().

The output of these systems can come in the form of a prediction or a list of recommended items. In the first case, the goal of the algorithm is approximating the rating a user would attribute to a yet unrated item, while the second type of output involves gathering the items that most likely would interest the user. Simple recommender systems that suggest items similar to the one being queried do not necessarily involve rating predictions, but it is common to have the list of recommended items based on the ratings the algorithms estimated the user would give to those items ().

Most recommender systems fall into one of four categories according to the filtering algorithm they use, that it, the strategy for generating predictions or selecting the top-N items: content-based filtering, demographic filtering, collaborative filtering, and hybrid filtering ().

Content-based filtering leverages characteristics of the content in order to generate the recommendations (Ricci *et al.*, 2011). One such algorithm might use the genres of watched movies in order to recommend new ones, while another might analyze the sound signature of a song to recommend similar ones, but, either way, all content-based systems establish a similarity between items as a basis for recommendations. Analogously, demographic filtering uses demographic data to establish a similarity between users and recommend items positively rated by similar people.

Collaborative filtering algorithms also recommend items that similar users liked, but, in this case, the similarity between users is based on past ratings and not demographic information (Ricci *et al.*, 2011). Hybrid filtering usually mix collaborative methods with either content-based or demographic filtering (Ricci *et al.*, 2011).

As with other knowledge-based systems, recommendation algorithms have quickly incorporated neural networks and other machine learning techniques over the past few years. Even though the implementation of YouTube's recommendation algorithm is a trade secret, it is known to gather enormous amounts of data about the user's interaction with the website and to require Google's own TPUs in order to be trained. It also involves two distinct steps: candidate generation (when the billions of videos available on the platform are quickly narrowed down to a few hundreds that might be relevant) and ranking (when the algorithm actually attempts to predict the score a user would implicitly give to the candidate videos) ().

Another relevant aspect of recommender systems that is well-exemplified by YouTube is the use of balancing factors such as novelty, dispersity, and stability (Zhao *et al.*, 2019). In the case of Google's video giant, there is a baked-in bias for recency, strongly favoring newer videos in detriment of older content (Zhao *et al.*, 2019).

From this kind of bias stems much debate: as recommender systems explode in popularity, so does research regarding its shortcommings. User radicalization and algorithmic bias (explicitly programmed or not) are hotly debated subjects in the literature.

1.3 Radicalization and Bias

Opinion polarization is far from a recent phenomenon, and social media is only the most recent communication medium where it can be detected and studied. An important question is whether it facilitates or attenuates polarization: anecdotal evidence might suggest that social network structures incentivize users to gather into antagonistic communities, but this could be a result of people simply being more likely to express their preferences online, not of some intrinsic property of social media.

One possible byproduct of polarization is radicalization. Despite not being entirely different phenomena, these concepts deserve distinct levels of attention. While polarization can be considered a natural part of democratic discourse, radicalization only happens when certain conditions are met. UNESCO defines radicalization as (Séraphin *et al.*, 2017):

- The individual person's search for fundamental meaning, origin and return to a root ideology;
- The individual as part of a group's adoption of a violent form of expansion of root ideologies and related oppositionist objectives;
- The polarization of the social space and the collective construction of a threatened ideal 'us' against 'them,' where the others are dehumanized by a process of scapegoating.

The third point is of special importance to the distinction between polarization and radicalization. The first might be a simple consequence of democratic disagreements between opposing parties, but the latter involves a dehumanization of the opposition, which can lead to extremism: radicalism so intense that the only effective strategy is physically exterminating the opposition.

Understanding how polarization might lead to radicalization (and, ultimately, to extremism) is, therefore, of paramount significance to cultivate healthy democracies, specially in the digital age. Since most social networks, as of this writing, are still poorly moderated, they allow users to be exposed to a plethora of viewpoints, from benign to insidious, possibly configuring a "pipeline of radicalization" through which regular users end up radicalized by coming into contact with extreme content ().

Of course this argument is still very much open for debate. Researchers have found evidences both for and against () the pipeline hypothesis and even proposed other means though which social media might help radicalize users (e.g. the supply and demand hypothesis ()). Despite all disagreements, one common point addressed by most research is the role of recommendation algorithms in serving users with radicalizing content.

Proponents of the pipeline hypothesis, for instance, argue that recommendation systems, aiming to maximize content consumption, suggest items that reinforce preconceived notions of the user and that play on fear and paranoia (). Content that appears urgent and leaves the user fearful (for their live, their community, or their identity) could be more engaging and, therefore, might be more susceptible to being considered as relevant by the algorithm.

Even if the pipeline hypothesis is correct, specifics of how much algorithms are to

blame for radicalization are still unknown and hard to pin down. Most research about the subject focuses on specific platforms (like Twitter and YouTube) and have severe limitations with regards to how much data those companies make available, not to mention the constant changes made to the algorithms over the years that might alter their radicalization properties. Definitive evidence for one theory or another must, therefore, apply to recommender systems in general and be predictive of how they work both in controlled and real life scenarios.

Closely related to user radicalization is the subject of algorithmic bias. YouTube, for example, has an explicit bias towards recency (), meaning that more recent videos get "boosted" by their recommendation algorithm. This is explicitly coded into the system, but there are also implicit biases, learned by watching user behavior.

Stoica () studied Instagram profiles before and after the implementation of their recommendation engine. They discovered that male users had a slight predilection for following other men, while women displayed no such preference and, as soon as the recommender system was deployed, engagement with profiles of male users skyrocketed even though they were the minority on the platform. The algorithm recognized and leveraged this so called differentiated homophily effectively, but we might question whether or not this should be the desired outcome of a good recommender system.

In social networks where recommendations are the source of most interations, such as YouTube (), the algorithm could go from being a mere reflection of user preferences to actively shaping user behavior. Hypothetically, a minoritary group of highly engaged users with strong self-reinforcing consumption habits could tip the scales of the algorithm and cause fringe content to be amplified; this is only one way through which a radicalization pipeline could spontaneously form in a social network.

1.4 Research Goals

As explained in the previous sections, social networks' recommendation algorithms might play a significant role in radicalizing users. This could be, at least in part, be a result of implicit and explicit biases in recommender systems.

This dissertation aims to explore the radicalization pipeline hypothesis and, more specifically, understand the mechanisms through which recommender systems can end up learning or developing biases. The research developed here revolves around the dynamical properties of recommender systems (i.e., the sequence of items suggested to an arbitrary user over time) and how feedback loops can create "amplification pipelines" inside these engines.

In short, the main motivator of this research is to test the pipeline hypothesis in a setting where recommendation algorithms learn dynamically.

1.5 Outline

...

Chapter 2

Literature review

There are three types of work that are relevant to the current topic: general literature about recommender systems, evidences of algorithmic bias, and methods of creating fairer recommendations. Since this area of study is still mostly unexplored, there is no consensus on whether social media recommender systems favor extremist content (or even whether they are actually deradicalisation agents), which means that many references used in this work might disagree amongst themselves.

2.1 Scientific literature

General literature about recommendation algorithms is abound. One of the most cited surveys was elaborated by Bobadilla *et al.* (2013), but works by He *et al.* (2016) (about interactive recommender systems), and by Kunaver and Požrl (2017) (about diversity in recommender systems) were also used in order to draw a complete panorama of the field.

Another relevant article, by Guy *et al.* (2010), is the landmark paper that inaugurates the usage of user data alongside labels to create a recommendation algorithm that is highly accurate and a staple of modern social networks. This essentially starts the usage of recommenders systems in social media.

When talking specifically about YouTube's recommendation algorithms, two papers deserve special attention. The first one, by Covington *et al.* (2016), marks YouTube's move towards the usage of deep neural networks to generate video recommendations. The authors describe a two-stage model that first generates a list of candidates and then ranks them, also reporting dramatic performance improvements. The second one, by Zhao *et al.* (2019), describing a more recent version of YouTube's recommendation algorithm, explores the Multi-gate Mixture-of-Experts technique to optimize recommendations for more than one ranking objective and the Wide & Deep framework to mitigate selection biases. The authors also make it clear that YouTube's recommender system has a strong bias towards more recent content instead of more traditional metrics.

Many authors have also explored how biases in recommendation engines might lead to user radicalization. Agarwal and Sureka (2015) developed an early example of a

technique to try and find extremist content on YouTube. Using advanced machine learning methods, the authors create a YouTube crawler that starts from a seed video and iteratively classifies featured channels and videos according to their potential extremism. A more recent example of this can be found in Tangherlini *et al.* (2020), where the authors propose a novel approach for identifying conspiracy theories online. By analyzing the narrative structure of a conspiracy theory (Pizzagate) and comparing it to an actual conspiracy (Bridgegate), they create a model that can guess whether a conspiratorial narrative is or not fabricated. According to their findings, a multi-domain nature and the presence of keystone nodes are signs that strongly indicate a conspiracy theory.

Besides just finding and identifying radicalizing content on YouTube, many authors have been concerned with studying the radicalization dynamics directly. Alfano *et al.* (2020), for example, claim to be "the first systematic, pre-registered attempt to establish whether and to what extent the recommender system tends to promote such [extremist] content." Cho *et al.* (2020) also attempt to understand how users can be radicalized by the algorithm. By experimentally manipulating user search/watch history, the authors concluded that algorithmically recommended content can reinforce a participant's political opinions.

In the same vein, Faddoul *et al.* (2020), after some high-profile cases of users being radicalized through YouTube videos, studied the efforts announced by the platform to curb the spread of conspiracy theories on the website. The paper aimed to verify this claim by developing both an emulation of YouTube's recommendation algorithm and a classifier that labeled whether a video is conspiratorial or not. The authors describe an overall decrease in the number of conspiracy recommendations, though not when weighing these recommendations by views.

Three papers that deserve a closer look are those that investigate how regular recommendation algorithms can learn covert biases in the users of a social network and amplify them to previously unimaginable rates. Stoica, Riederer, *et al.* (2018) explore the existence of an "algorithmic glass ceiling" and introduces the concept of differentiated homophily. The authors experiment on a Instagram dataset before and after the introduction of algorithmic recommendations and discover that, even though most of that network's users were female, the most followed profiles were male. They explain this phenomenon by postulating that the algorithm learns biases in the population, that is, male preference for male profiles (which doesn't happens for females and thus characterizes an asymmetric—differentiated—homophily), and ends up enhancing this effect. Stoica and Chaintreau (2019), building on top of their previous work, create a proposal for new recommender systems that take differentiated homophily into account in order to reduce the "glass ceiling" effect observed in non-corrected recommendation algorithms. The work focuses on the theoretical description of the algorithm, but also attempts to validate its hypothesis in real world data. Stoica (2020), in their most recent paper, show that the most commonly used metrics in recommender systems "exacerbate disparity between different communities" because they reinforce homophilic behavior of the network. This has profound implications, since these algorithms might further suppress already minoritary viewpoints without being explicitly programmed to do so.

Like the aforementioned articles, Matakos *et al.* (2020) also propose a novel recommen-

dation algorithm that tries to strike a balance between information spread and ensuring that the users are exposed to diverse viewpoints. The authors show that this goal is important if we want to foster healthy online debate, and that the algorithm is efficient and scalable with a minor approximation. One possible inspiration for these papers might be one by Su *et al.* (2016) that studyed the network structure of Twitter before and after the introduction of algorithmic recommendations ("Who to Follow"). The authors of the paper discovered that all users benefitted recommendations, but that users with already popular profiles benefitted even more, effectively changing the network structure and dynamics. Caton and Haas (2020) have recently compiled other valuable information on fairness in machine learning into a survey.

Because of data limitations, there still are few studies that investigate how recommendation algorithms work dinamicaly, over time. Burke (2010) point out that most methods for evaluating recommender systems are static, that is, involve static snapshots of user and item data. The authors propose a novel evaluation technique that helps provide insight into the evolution of recommendation behavior: the "temporal leave-one-out" approach. A more recent example of this approach was developed by Roth *et al.* (2020). Their paper delves into the confinement dynamics possibly fostered by YouTube's recommendation algorithm. The authors create, from a diverse set of seed videos, a graph of the videos iteratively recommended by YouTube and, from this, study whether there were created "filter bubbles". They find that indeed YouTubes recommendations are prone to confinement dynamics be it topological, topical or temporal.

Even more recently, Yao *et al.* (2021) propose an approach for measuring recommender system bias based on simulated users. Even though this work focuses only on bias towards popular content, it is of particular importance because it was written by researchers from Google itself. Some years before, Dash *et al.* (2019) also proposed a framework for auditing recommender systems based on its network of users. Another contribution of their work is a novel quantifications of diversity.

A different approach to understanding biases in recommendation algorithms range from analyzing similarity metrics to developing theoretical bounded confidence models. Giller (2012) goes with the first strategy, and identifies certain aspects of cosine similarity that are often overlooked. Starting from simple theorems regarding the density of n-dimensional spheres, the author concludes that the expected cosine similarity between random bitstreams might be significantly different from the average. This is noteworthy because many recommendation algorithms use cosine similarity in order to determine the similarity between two items to recommend. Sîrbu *et al.* (2019) go with the latter, providing an interesting theoretical model of how inherent biases in algorithmic recommendations might highten opinion polarization. Using a bounded confidence model, the authors propose the addition of a γ term that represents the odds of an algorithm recommending content that differs from that of a user.

Some recent papers also try to understand how YouTube might be favoring right-wing and fascist content in specific, as opposed to trying to prove a more general (and possibly less tractable) claim. Hosseinmardi *et al.* (2020) find evidence via a longitudinal study that there exists "a small but growing echo chamber of far-right content consumption" on YouTube. According to their research, these users are more engaged than other, with

YouTube generally accounting for a larger share of their online news diet than the average. The authors, however, find no evidence of this phenomenon being due to recommendations. A popular article in the field, by M. H. Ribeiro *et al.* (2020), explored the radicalization pipeline hypothesis of algorithmic enabled radicalization. The authors collect huge amounts of YouTube comment data over time, and determine a significant migration of users from "lighter" content towards more extreme videos. This does not prove that the pipeline exists, but is a strong argument for its existence.

Twitter was also found to consistently favor right-wing content. Huszár *et al.* (2021) conducted a "long-running, massive-scale randomized experiment" across 7 countries in order investigate the effects of algorithmic personalization on users' feeds and, according to their results, "mainstream political right enjoys higher algorithmic amplification than the mainstream political left".

Finally, feedback loops are of special interest to this discussion. Caused by the inevitable fact that recommender systems must learn from users' reactions to its own recommendations, they are a widely believed to be a powerful engine of bias amplification and are discussed at length in the literature. Already in the last decade, Sinha *et al.* (2017) investigated the viability of identifying items affected by these feedback loops and attempted to created a method of deconvolving them. More recently, Jiang *et al.* (2019) explored what they called "degenerate feedback loops" and their capability of creating echo chambers, going as far as proposing a novel approach of slowing down this tendency towards degeneracy. In a related study, Mansoury *et al.* (2020) explored how recommender systems amplify already popular content, but, more importantly, how this tendency might reduce content diversity and cause users' tastes to shift over time. Depending on what a systems values (recency, virality, controversy, engagement), this type of feedback loop could possibly amplify not "popular" content, but divisive and extremist content.

A minortiy of papers tries to disprove the hypothesis that social networks in general, and YouTube in specific, have a radicalizing tendency. Munger and Phillips (2020) published a controversial article that postulates a new model for YouTube radicalization. According to the authors, YouTube's algorithm is not to blame, the users themselves are looking for extreme content and the recommender system only supplies them. Its methods were highly questioned by the community and is currently the only paper that spouses the supply and demand hypothesis. Ledwich and Zaitsev (2019) also wrote a highly controversial paper where its authors claim to have found evidence to support the hypothesis that YouTube's recommendation algorithm favors mainstream and left-leaning channels instead of right-wing ones. They categorize almost 800 channels into groups of similar political leaning and analyze recommentations between each group, finding that YouTube might actually discourage users from viewing radicalizing content. Most researchers though do not support the methods employed by these two articles. In an even earlier study on news recommendations of a major Dutch newspaper, Möller *et al.* (2018) claim that recommenders systems had no significant impact on content diversity.

Even with a quickly growing body of research, further studies are needed in order to shed more light into the inner workings of how recommendation algorithms are used by social networks. Articles like the ones described in this chapter are of utter importance to this task, but generalist studies that are able to capture dynamics common to all or most

recommender systems are still nonexistent.

2.2 Journalistic efforts

Since this field of study is still in its infancy, many relevant sources are not scientific in nature. Journalism, specially when investigative in nature, is a valuable ally when trying to understand what is happening behind the curtains of social platforms.

Some examples of journalistic endeavors that inform and guide scientific research include, but are not limited to, a series by Lecher and Yin (2022) on how different are Americans' Facebook feeds, a report (in Portuguese) by P. V. Ribeiro (2021) on how the far-right is still able to cheat YouTube's attempts at curbing extremist content, and a whistleblower's account to Wong (2021) of how Facebook's executives resist on restricting fake engagement that is able to distort global politics.

Chapter 3

Static Analysis

As discussed in the previous chapters, understanding how social networks recommend content to users is central to the debate around political polarization and radicalization. There are many ways of exploring recommender systems without examining their code (), from simulating their behavior after careful observation to directly collecting recommendation data, but most of them allow us to examine only one perspective of the algorithm at a time. This means that studying a social network's recommendation technique has inherent limitations.

Most of the algorithms currently employed by social media companies are trade secrets. They are also subject to constant experimentation and tuning (), which might render worthless any research performed before an update to the algorithm, no matter how carefull the design of the study was. YouTube, for example, currently has over 2 billion monthly logged-in users (), but it makes no significant effort to clarify changes made to the algorithm or even whether they fulfill their promisses of reducing user exposure to radicalizing content. With more than 500 hours of content being uploaded every minute (), if 1% of all videos can be considered radicalizing and the algorithm can detect 99% of them, that still leaves over 25.000 hours of brand new extremist content free to spread on the platform every year. This goes to show that, in the scale that these companies operate, even a small fraction of content might still be enough to influence the overall recommendations made by the algorithm. It is also worth noting that most of these platforms' efforts are concentrated in their parent countries (usually the United States), so, even if they actually try and remove the offending content, most of the non-English-speaking world would still not be impacted by their policy changes.

This leads right into the goal of the present dissertation. With our experiments we aim to make a tangible contribution to the field of recommender systems, specifically how their design might (or might not) foster confinement dynamics in the "phase space" of recommendations. If the main hypothesis is confirmed, this could mean that even a relatively small fraction of the content can tip the algorithm in its favor, amplifying their message, creating "filter bubbles", and possibly sending users on a radicalization spiral if that topic is related to politics or other contentious subjects.

3.1 Datasets

Before discussing any experiment, it is necessary to introduce the datasets used to train the models. The main dataset explored in this dissertation is MovieLens (Harper and Konstan, 2015), a well-known set of movie reviews that has been featured in many recommender system tutorials and papers for the past few years. The full dataset, with 27,000,000 ratings applied to 58,000 movies, was enriched by Banik (2017) with information about the movies' credits, metadata, keywords, and links. In the end, because of technical limitations, the dataset used in this chapter was sampled until 30,689 movies were left; this allowed for faster experimentation and simpler plots.

The second dataset, used for validation purposes only, was the Book-Crossing Dataset (Ziegler, 2004). Just like the enriched MovieLens, this dataset contained entries for ratings (1,149,780) applied by users (278,858) to items (271,379 books), and information about these items like title, author, publisher, etc.

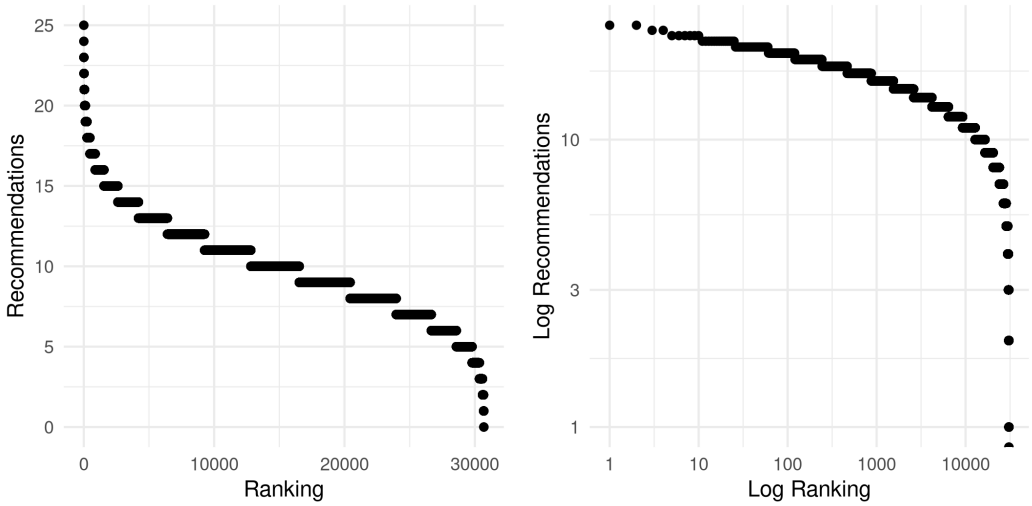
3.2 Experiments

Some preliminary were conducted in order to gather some evidence in favor or against the main hypothesis being tested. In total, fifteen different recommendation models were trained and analyzed, with each plot below representing one of these models. For simplicity's sake, even though all models represented here are non-parametric, we still say they were "trained" because the datasets used to generate recommendations are different.

The goal of these experiments was trying to identify if even a simple recommendation algorithm could demonstrate some sort of bias towards a subset of the items being recommended. More specifically, given an algorithm that cannot be influenced by users' personal preferences, would the resulting recommender system favor some kind of item? Excluding user information is important because, as demonstrated by Stoica, Riederer, *et al.* (2018), users might have their own biases and these could get transferred on to the model; the objective here is understanding the algorithm by itself without external influences.

To evaluate the recommendation models, at least qualitatively, the authors plotted their "recommendation profiles": a summary of how many times an arbitrary item is recommended overall. To create this profile, the algorithm is asked to return the top-n most similar items to the input according to its internal similarity metric, and this process is repeated for every item in the dataset. The recommendation profile of the model is the number of times each item showed up in the top-n most similar items of the whole dataset.

For example, a recommendation algorithm trained on the present version of the MovieLens dataset would generate a list of n movies for each input, creating a list of $30,689 \times n$ movies. After this computationally intensive calculation, the occurrence of each ID would be counted and ranked accordingly to facilitate interpretation of results, that is, the movie ranked number 1 would be the movie featured the most times in the set of all recommendations, and so forth for every other rank. From now on, n will be fixed to 10.



(a) *Trivial recommendations.* **(b)** *Log-log plot.*

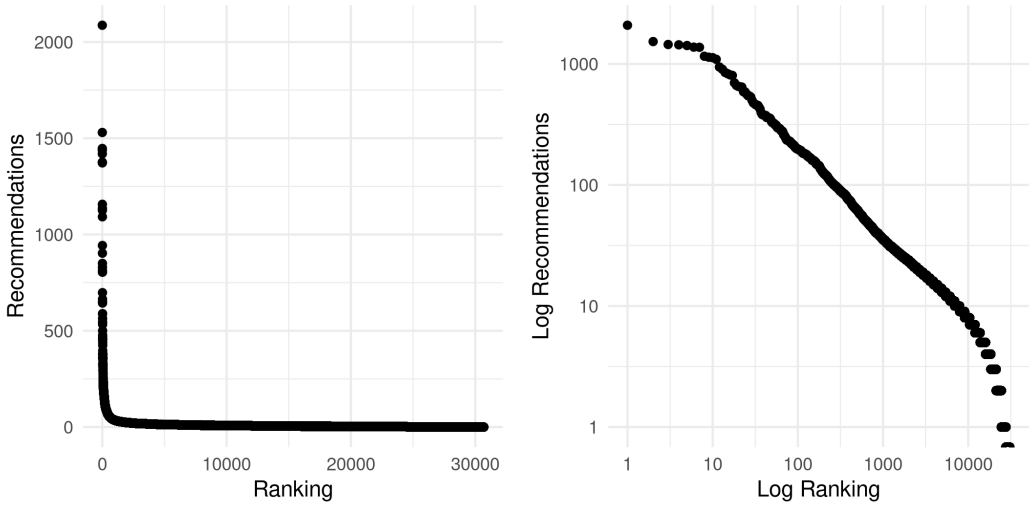
Figure 3.1: *Recommendation profile for the trivial model (a) and log-log plot (b).*

The baseline against which all models will be compared can be seen in Figure 3.1, the so called "trivial model". This model is a simple sampler that returns n movies at random when asked for a recommendation and, thus, its recommendation profile averaged, evidently, n, with an appearance very similar to that of the CDF of the normal distribution. The "most recommended" movie appeared 25 times in the final list, while the "least recommended" movie did not appear at all. The log-log plot (Figure 3.1b), despite seeming out of place, will make sense when the second model is presented.

Besides the trivial model, the simplest model that excludes user information is the content-based recommender. In the real world this is an algorithm that is able to identify similar items based on their metadata (description, tags, etc.) and suggest the closest items to the one being purchased or viewed. A straightforward way of building such an algorithm is creating a vector representation of each item and then using a similarity metric to recommend the items most similar to the one in question. The chosen similarity metric was cosine similarity because of its simplicity, robustness, and ubiquity (Sarwar *et al.*, 2001).

The main non-trivial model used in this study was the one that simply generated vector representations for the full MovieLens dataset, without any modifications (which is why it will henceforth be referred to as the "vanilla" model). The metadata for each movie was made up of its keywords, main cast, director, and genres. The vector transformation was very simple, with each position representing one of the words of the corpus, and each element indicating how many times that word appeared in the metadata for that movie. When the recommendation for a movie was requested, the algorithm measured the cosine similarity between it and every other movie, returning the IDs belonging to the top $n = 10$ most similar vectors.

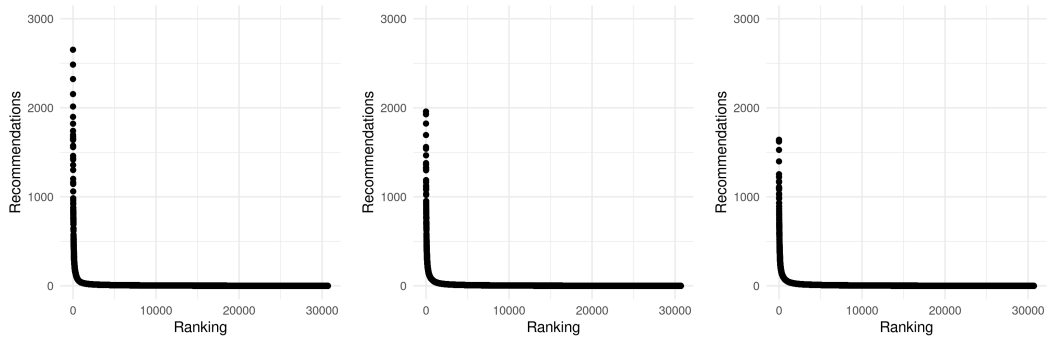
The recommendation profile for the vanilla model can be seen if Figure 3.2a. Compared to Figure 3.1a, the same visualization for the trivial model, the distribution of the vanilla model differs immensely: the movie ranked number 1 appeared more than 2000 times in



(a) *Rec. based on movie metadata.* **(b)** *Log-log plot.*

Figure 3.2: *Recommendation profile for the vanilla model (a) and log-log plot (b).*

the full list of recommendations, with an almost exponential decrease in the number of appearances from then on. The log-log plots of both models (Figure 3.1b and Figure 3.2b, respectively), makes it clear that the vanilla model is close to exponential, while its trivial counterpart is (evidently) normal.

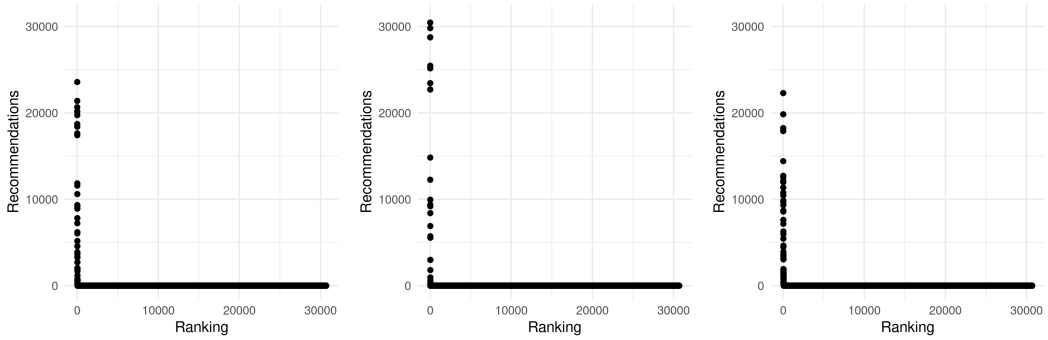


(a) *Cutoff* $k = 2$. **(b)** *Cutoff* $k = 5$. **(c)** *Cutoff* $k = 8$.

Figure 3.3: *Recommendation profile for cutoff* $k = 2$ *(a),* 5 *(b), and* 8 *(c).*

A potential explanation for the difference between trivial and vanilla could reside in the least used terms in the metadata: movies whose metadata share rare words might have been recommended less frequently than movies whose metadata is not so unique. To test this hypothesis, a cutoff point was created for words to be included in the vector representation of the movies. Three cutoff points were tested where only words with an absolute frequency larger than or equal to k, $k = 2, 5, 8$, could be included in the vector representations. The results can be seen in Figure 3.3 and, aside from variations in the y-intercept, all plots are qualitatively very similar to Figure 3.2a, indicating that rare words probably are not to blame for the exponential-like decay.

The second validation experiment involved attempting to use other distance metrics instead of cosine similarity (Ricci *et al.*, 2011), since that could also be a source of the

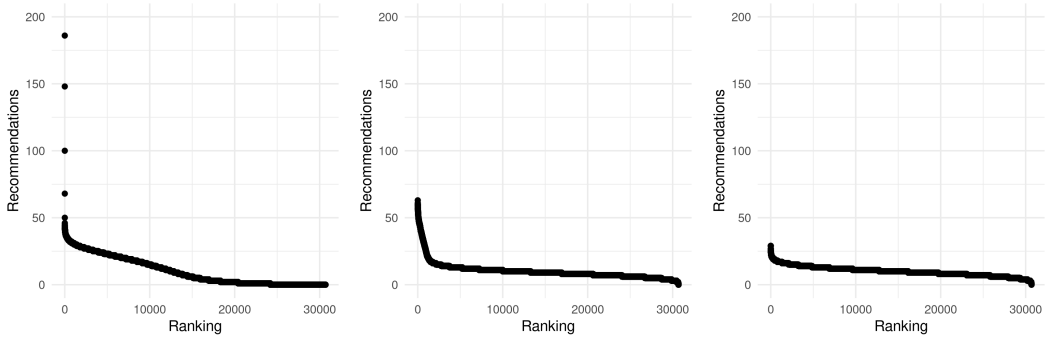


(a) *Cosine distance.* **(b)** *Euclidean distance.* **(c)** *Manhattan distance.*

Figure 3.4: *Recommendation profile for cosine (a), euclidean (b), and manhattan (c) distances.*

strange behavior of the recommendation profile. The goal here was verifying whether other metrics could do a better job at not creating a subset of movies that ended up exponentially more recommended than the rest. As attested by Figure 3.4, this was not the case.

At this point it is safe to say that the type of decay seen in recommendation frequencies up until now is not spurious and must have a clear cause. A hypothesis that is later mostly confirmed involves the average number of non-zero elements in the vector representation of the movies: the sparser the vectors, the higher the odds of the recommendation curve displaying a steep left-hand side. This would be almost equivalent to creating an "inverse cutoff point", removing common words from the vector representations.



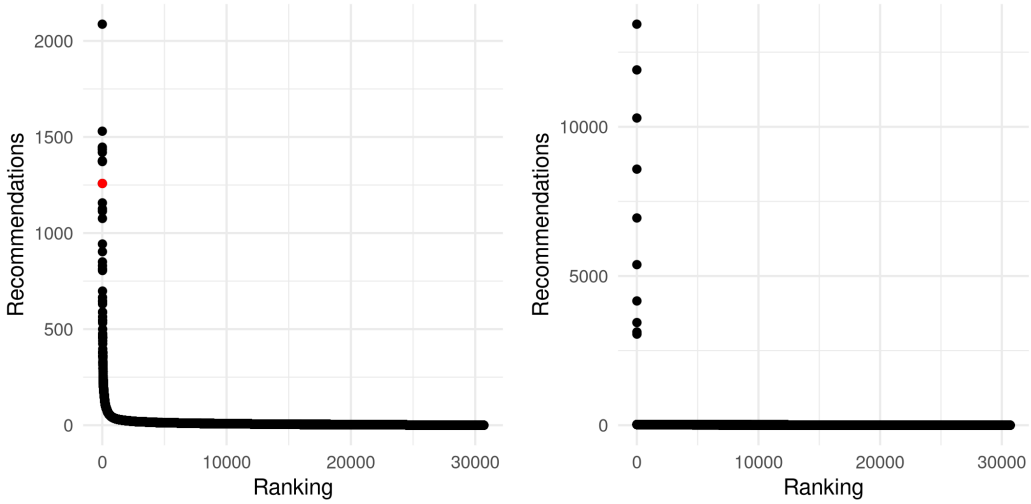
(a) $P(w_i) = 1 \times \overline{P(w)}$. **(b)** $P(w_i) = 10 \times \overline{P(w)}$. **(c)** $P(w_i) = 100 \times \overline{P(w)}$.

Figure 3.5: *Recommendation profile of samples with* $P(w_i) = C \times \overline{P(w)}$, $C = 1$ *(a),* 10 *(b), and* 100 *(c).*

To test whether this hypothesis held water, "random" vector representations were created. These representations were based on fictional metadata that were comprised of words sampled at random from the full corpus of movies; the probability that word w_i occurred in the metadata of a fictional movie was equal to the average probability that any word would occur in some arbitrary metadata ($\overline{P(w)}$) times a costant C, $C = 1, 10, 100$. More simply, the probability of an element being non-zero in a random vector representation was the average probability that an arbitrary element of the vanilla representations was non-zero times C. The constant was added as a way to create less sparse vectors and allow for comparisons between different inverse cutoff points.

Figure 3.5 displays the recommendation profiles for each different C. Concretely, the

figures are equivalent to creating random metadata for the movies where the probability of a word occurring was approximately 1.54×10^{-4}, 1.54×10^{-3}, and 1.54×10^{-2} respectively and then training the recommendation models. The results do support the aforementioned hypothesis since less sparse vectors indeed generated less exponential decays.



(a) *Vanilla model with artificial movie in red.* **(b)** *Model with short vector representations.*

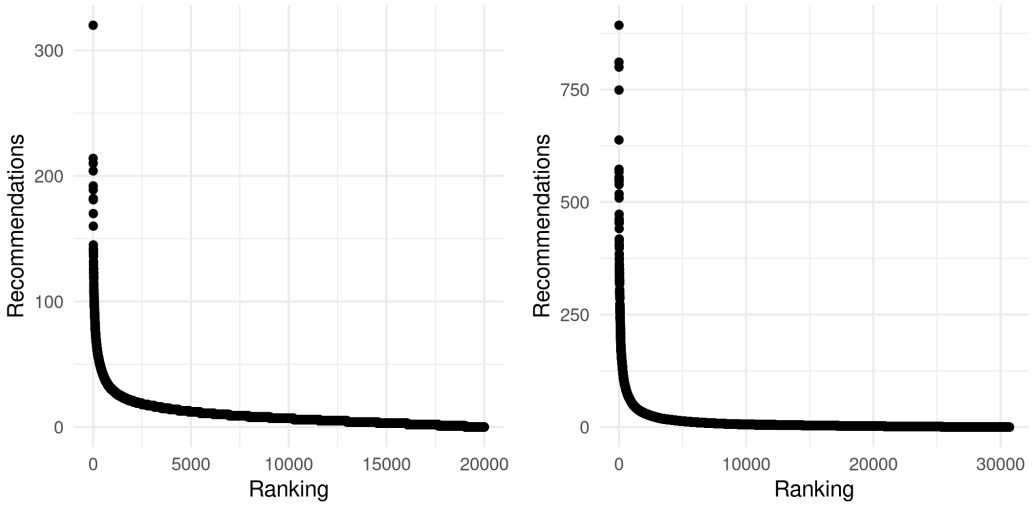
Figure 3.6: *Recommendation profile for artificial movie (a) and short vector representations (b).*

After the previous experiments, sanity checks were needed in order to guarantee that the previous hypothesis was able generalize. The first check should verify whether an artificial movie created as a combination of the metadata from other movies favored by the recommendation algorithm would also be favored, while the second should check whether shorter vectors would change the decay already observed despite being as sparse as their longer counterparts.

Figure 3.6 showcases the two sanity checks. Figure 3.6a was a model trained with the vanilla dataset with the addition of the movie highlighted in red. As expected, this movie also showed up in the top-recommended subset. Figure 3.6b comes from a model trained on randomly generated vector representations in a similar fashion to the ones in Figure 3.5, except each vector could only have 15,000 elements instead of 55,681 (as with the vanilla model). The patter observed before persisted, meaning that the hypothesis still stood.

The last two models were considered the confirmations of the hypothesis that (at least for this kind of recommendation systems) a subset of items was always much more recommended than the rest as long as the data was sparse. Figure 3.7a represents the same recommendation algorithm applied to another dataset, the Book-Crossing dataset. Figure 3.7b contains the results of the model applied to another set of random vector representations, this time with the probability of each element being non-zero respecting the marginal distributions of the vanilla dataset. Again, the exponential decay pattern persisted, only slightly less pronounced in the Book-Crossing case.

The analysis up until now has been static, that is, the recommendation model is trained and applyied to every movie in the dataset. There is no interaction with users



(a) *Recommendation profile for book dataset.* **(b)** *Random simularion of vanilla.*

Figure 3.7: *Recommendation profile for book dataset (a) and random simulation of vanilla (b).*

and no opportunity to evolve over time. The next chapter adresses this point by using Google's newly released TensorFlow Recommenders library (*TensorFlow Recommenders* 2021) to gather data about what happens to a system's recommendations as users follow its suggestions. Employing a deep learning model that is able to improve over time is a significant departure from the content-based models presented here and, if a similar recommendation profile can also be detected for multi-criteria recommender systems on dynamic scenarios, then the hypothesis ventilated in the section above would become even more plausible.

Chapter 4

Dynamic Analysis

Following the static analysis, it became clear that a dynamic analysis would be of the utmost importance. Understanding how the recommendation model responds to users reinforcing its internal biases, like the ones already detected, could potentially lead to a better understanding of how these systems favor certain kinds of content.

In order for this analysis to be more true to reality, we implemented a simple recommendation algorithm using TensorFlow Recommenders (), a library for machine learning developed by Google for use with its TensorFlow () framework. This means that, even though our model is deliberately bare-bones, it conforms to industry-standard technology and practices.

The choice to use a simple recommendation algorithm instead of a more complex one was twofold: first, we didn't want to use a model that could introduce many confounding parameters to the analysis (e.g. hyperparameters, hardware requirements, etc.), and second, we wanted to study a baseline that could, in the future, be used as a comparison point for more complex algorithms.

The goal of this analysis is to gather data on how the recommendation system behaves over time. As will be explained in the next sections, is to understand what happens to the recommendation profile of the algorithm as it interacts with itself via users that follow the generated suggestions.

The expectation is that the recommendation profile will grow ever more steep, which is a reasonable guess; if the users reinforce the beliefs of the algorithm, then is stands to reason that it will recommend popular movies with more and more frequently, to more and more users. How much more frequently, however, is the true question.

For the sake of clarity, let's imagine two users with very distinct preferences: Alice, who enjoys adventure movies, and Bob, who enjoys horror movies. In principle, the algorithm should have very different recommendations for both of them and, were they to follow them, their custom suggestions should grow increasingly different. At the end of this experiment, users like Alice would all be recommended the same movies, and users like Bob would have their own set of very popular films; we should expect, therefore, a multimodal distribution of the recommendation frequencies, with "typical" adventure movies and

"typical" horror movies being much more popular than comedy, for example.

However, if the final recommendation profile looked like what was showcased in the previous chapter, i.e. a very small subset of movies being recommended to most users, then we could infer that the system devolved into a degenerate feedback loop, ignoring personal preferences and distinctions between films.

4.1 Datasets

For the dynamic experiments, we kept using the Movielens dataset (). This time, however, we used the full "1M" dataset instead of sampling movies from the larger "25M" version. Given that we wanted our dynamic analysis to be conducted in a realistic scenario, we decided that it would be better not to change the data. This whole experiment will, therefore, use a version of the dataset commonly used for machine learning benchmarks with no alteration whatsoever.

The 1M dataset contains 1,000,209 ratings of almost 4000 movies made by over 6000 anonymous MovieLens users who joined the platform in 2000. In this particular version, each user has made at least 20 ratings. There are 4 columns available:

- `UserID`: Unique user identifier, ranging from 1 to 6040.
- `MovieID`: Unique movie identifier, ranging from 1 to 3952.
- `Rating`: Movie rating according to user, from 0 to 5 stars.
- `Timestamp`: When the user made the rating, in seconds since the epoch.

A second, auxiliary, dataset was also used to enrich the main one. "Movies" contains extra information about the movies in 1M, which allowed us to add more variables to the recommendation system. This new dataset has 3 columns:

- `MovieID`: Unique user identifier, ranging from 1 to 6040.
- `Title`: Title of the movie, as provided by IMDB.
- `Genres`: Pipe-separated string with all applicable genres.

The other accompanying dataset, "Users", has not been used for the purposes of this analysis. The reasoning behind this decision will be explained in greater detail in the next section.

4.2 Experiments

The dynamic experiment starts in a manner much similar to the static experiment. The full MovieLens dataset is fed as training data to a recommendation system in order to get it ready for giving suggestions to users. As explained in the opening section of this chapter, we chose a simple algorithm in order to reduce the number of possible interferences architecture could have on our analysis.

The chosen recommendation algorithm was a basic ranking model described in using TensorFlow Recommenders (). It is composed of multiple stacked dense layers and uses mean squared error as its loss function. The main class in the model is reproduced below, and the full algorithm is listed in Appendix ??.

```
class MovielensModel(tfrs.models.Model):

  def __init__(self):
    super().__init__()
    self.ranking_model: tf.keras.Model = RankingModel()
    self.task: tf.keras.layers.Layer = tfrs.tasks.Ranking(
      loss = tf.keras.losses.MeanSquaredError(),
      metrics=[tf.keras.metrics.RootMeanSquaredError()]
    )

  def call(self, features: Dict[str, tf.Tensor]) -> tf.Tensor:
    return self.ranking_model(
        (features["user_id"], features["movie_title"]))

  def compute_loss(self, features: Dict[Text, tf.Tensor],
                   training=False) -> tf.Tensor:
    labels = features.pop("user_rating")

    rating_predictions = self(features)

    # The task computes the loss and the metrics.
    return self.task(labels=labels, predictions=rating_predictions)
```

In the first step of the experiment, we trained the recommendation model using Movielens' 1M ratings dataset, which we will refer to as `ratings0` from now on. All available data was used and, in the end, we achieved a root mean squared error (RMSE) of 0.92; this result is similar to TFRS' deep & cross network () results when trained on the same data. Once `model0` was ready for making recommendations, we applied it to every possible user-movie pairing, generating a complete matrix of predicted ratings called `predictions0`.

In an environment like YouTube's recommendations sidebar, the user is presented with a few items that the algorithm thinks they would like, and then they can either ignore the sidebar or select one of the options to watch. Since our goal was to explore what would happen when the recommendation system entered a feedback loop, we picked one movie at random from each user's 10 best-ranked entries.

This set of well-ranked movies was our way of simulating thousands of users simultaneously approving of the algorithms recommendations and selecting one option from their sidebars. The last step involved removing the oldest rating of each each user from `ratings0` and appending these these selections to the dataset in order to create `ratings1`. The full data flow is illustrated in Figure 4.1

A feedback loop, however, can't be created from a single iteration. For this reason,

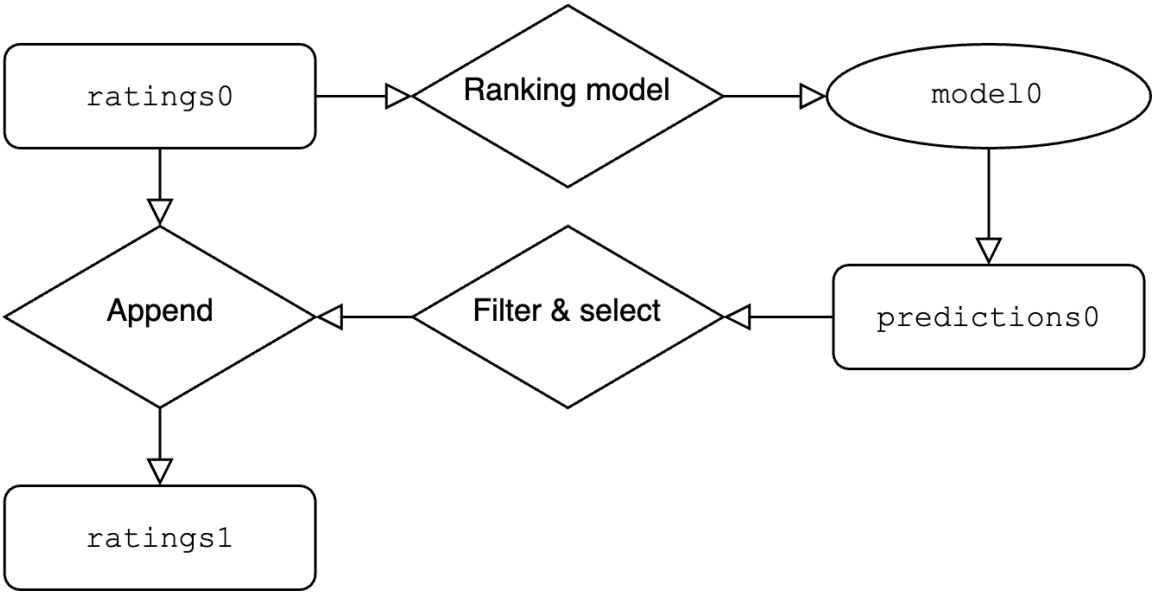


Figure 4.1: *Data flow diagram.*

the process just described was classified as the "zeroth" iteration in the process (as it involved `ratings0`). The "first" iteration began with `ratings1`, trained `model1`, generated `predictions1`, and ended with `ratings2`. We repeated this process until we got to `ratings4`, totaling 5 ratings datasets (one original and 4 derived through ranking models).

A simplified version of the R code that created `ratingsN+1` from `ratingsN` can be found below. The omitted functions served mostly auxiliary purposes which made the datasets conform to TensorFlow's format expectations.

```
predictions |>
  group_by(user_id) |>
  slice_max(prediction, n = 10) |>
  slice_sample(n = 1) |>
  ungroup() |>
  bind_rows(old_ratings) |>
  group_by(user_id) |>
  slice_max(timestamp, n = -1, with_ties = FALSE) |>
  ungroup()
```

With these new datasets we were able to analyze the differences between distinct generations of models and understand exactly how the positive feedback loop influenced the last iteration.

A series of analysis were conducted using these datasets as sources. We found that, with each iteration, the recommendation profile got steeper and steeper, that is, a few popular items got more recommended while the rest fell into disfavor; this was predictably more noticeable in movies with higher average ratings. In fact, a small set of around 20 movies were the only ones that consistently rose in popularity with new iterations.

In order to understand how this "feedback loop" was developing, we fit multiple regression models on our generated data. In the expression below, `pop` represents the popularity of a movie, `t` represents the time i.e. the iteration from 0 to 4, `genre` represents the genre of a movie, and `movie_id` is the ID of a movie. Note that the genre was not used when training the recommendation models described above, but it turned out that this feature could explain a lot of the models' outputs.

```
  Family: nbinom2  ( log )
Formula:          pop ~ t * genre * rating + (1 | movie_id)
Data: features

     AIC      BIC   logLik deviance df.resid
 83863.8  84370.3 -41865.9  83731.8    15844

Random effects:

Conditional model:
 Groups   Name        Variance Std.Dev.
 movie_id (Intercept) 1.364    1.168
Number of obs: 15910, groups:  movie_id, 3182

Dispersion parameter for nbinom2 family ():   49

Conditional model:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.108714   0.272464  -0.399 0.689891
t                         -0.210272   0.021153  -9.941  < 2e-16 ***
genreAdventure             0.112129   0.574197   0.195 0.845174
genreAnimation            -2.286362   0.773873  -2.954 0.003132 **
genreChildren's            0.855196   0.755919   1.131 0.257915
genreComedy               -0.108159   0.352431  -0.307 0.758925
genreCrime                -3.590470   0.850227  -4.223 2.41e-05 ***
genreDocumentary          -1.659325   1.508329  -1.100 0.271285
genreDrama                -1.219969   0.409312  -2.981 0.002877 **
genreFilm-Noir           -16.633103   2.371128  -7.015 2.30e-12 ***
genreHorror                0.512807   0.443025   1.158 0.247062
genreMusical              -4.259252   2.088483  -2.039 0.041410 *
genreMystery              -4.646920   1.461852  -3.179 0.001479 **
genreRomance              -2.118789   1.856289  -1.141 0.253699
genreSci-Fi                0.363380   1.044431   0.348 0.727899
genreThriller              1.382908   0.855392   1.617 0.105944
genreWestern              -3.766758   2.108848  -1.786 0.074072 .
rating                     0.957533   0.085468  11.203  < 2e-16 ***
t:genreAdventure           0.161327   0.053403   3.021 0.002520 **
t:genreAnimation          -0.392012   0.067698  -5.791 7.01e-09 ***
t:genreChildren's          0.116028   0.066946   1.733 0.083068 .
t:genreComedy              0.119502   0.030504   3.918 8.94e-05 ***
```

```
t:genreCrime                -0.146821   0.085503  -1.717 0.085952 .
t:genreDocumentary           0.329357   0.195776   1.682 0.092507 .
t:genreDrama                 0.005393   0.038739   0.139 0.889287
t:genreFilm-Noir            -0.730742   0.227275  -3.215 0.001303 **
t:genreHorror                0.148203   0.041813   3.544 0.000393 ***
t:genreMusical               0.265118   0.243060   1.091 0.275383
t:genreMystery              -1.150785   0.123034  -9.353  < 2e-16 ***
t:genreRomance               0.064921   0.227735   0.285 0.775590
t:genreSci-Fi               -0.311108   0.079779  -3.900 9.63e-05 ***
t:genreThriller              0.134758   0.077077   1.748 0.080403 .
t:genreWestern               0.184582   0.216581   0.852 0.394073
t:rating                     0.029594   0.006273   4.717 2.39e-06 ***
genreAdventure:rating       -0.209253   0.179840  -1.164 0.244606
genreAnimation:rating        0.594327   0.226301   2.626 0.008633 **
genreChildren's:rating      -0.473591   0.247951  -1.910 0.056131 .
genreComedy:rating          -0.194909   0.109223  -1.785 0.074341 .
genreCrime:rating            0.736375   0.243050   3.030 0.002448 **
genreDocumentary:rating     -0.113489   0.399299  -0.284 0.776242
genreDrama:rating           -0.031020   0.120957  -0.256 0.797601
genreFilm-Noir:rating        3.958891   0.595316   6.650 2.93e-11 ***
genreHorror:rating          -0.422452   0.151685  -2.785 0.005352 **
genreMusical:rating          0.944897   0.569196   1.660 0.096903 .
genreMystery:rating          1.092607   0.409047   2.671 0.007560 **
genreRomance:rating          0.013712   0.548422   0.025 0.980053
genreSci-Fi:rating          -0.423339   0.324094  -1.306 0.191477
genreThriller:rating        -0.729129   0.256016  -2.848 0.004400 **
genreWestern:rating          0.725673   0.578412   1.255 0.209625
t:genreAdventure:rating     -0.053250   0.015828  -3.364 0.000768 ***
t:genreAnimation:rating      0.110189   0.018603   5.923 3.16e-09 ***
t:genreChildren's:rating    -0.039499   0.020922  -1.888 0.059031 .
t:genreComedy:rating        -0.039790   0.008913  -4.464 8.04e-06 ***
t:genreCrime:rating          0.038460   0.022654   1.698 0.089557 .
t:genreDocumentary:rating   -0.088874   0.050610  -1.756 0.079076 .
t:genreDrama:rating         -0.006800   0.010754  -0.632 0.527202
t:genreFilm-Noir:rating      0.183567   0.054187   3.388 0.000705 ***
t:genreHorror:rating        -0.052874   0.013574  -3.895 9.81e-05 ***
t:genreMusical:rating       -0.082105   0.063538  -1.292 0.196285
t:genreMystery:rating        0.335177   0.032553  10.296  < 2e-16 ***
t:genreRomance:rating       -0.018634   0.064877  -0.287 0.773939
t:genreSci-Fi:rating         0.108210   0.023442   4.616 3.91e-06 ***
t:genreThriller:rating      -0.044310   0.022584  -1.962 0.049758 *
t:genreWestern:rating       -0.055461   0.056878  -0.975 0.329521
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

References

[Agarwal and Sureka 2015] Swati Agarwal and Ashish Sureka. "Topic-Specific YouTube Crawling to Detect Online Radicalization". In: *Databases in Networked Information Systems*. Ed. by Wanming Chu, Shinji Kikuchi, and Subhash Bhalla. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 133–151. ISBN: 978-3-319-16313-0. DOI: 10.1007/978-3-319-16313-0_10 (cit. on p. 7).

[Alfano *et al.* 2020] Mark Alfano, Amir Ebrahimi Fard, J. Adam Carter, Peter Clutton, and Colin Klein. "Technologically scaffolded atypical cognition: the case of YouTube's recommender system". In: *Synthese* (June 9, 2020). ISSN: 1573-0964. DOI: 10.1007/s11229-020-02724-x. URL: https://doi.org/10.1007/s11229-020-02724-x (visited on 12/02/2020) (cit. on p. 8).

[Banik 2017] Rounak Banik. *The Movies Dataset*. Nov. 10, 2017. URL: https://kaggle.com/rounakbanik/the-movies-dataset (visited on 03/01/2021) (cit. on p. 14).

[Bobadilla *et al.* 2013] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. "Recommender systems survey". In: *Knowledge-Based Systems* 46 (July 1, 2013), pp. 109–132. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2013.03.012. URL: http://www.sciencedirect.com/science/article/pii/S0950705113001044 (visited on 10/29/2020) (cit. on p. 7).

[Burke 2010] Robin Burke. "Evaluating the dynamic properties of recommendation algorithms". In: *Proceedings of the fourth ACM conference on Recommender systems*. RecSys '10. New York, NY, USA: Association for Computing Machinery, Sept. 26, 2010, pp. 225–228. ISBN: 978-1-60558-906-0. DOI: 10.1145/1864708.1864753. URL: https://doi.org/10.1145/1864708.1864753 (visited on 10/29/2020) (cit. on p. 9).

[Caton and Haas 2020] Simon Caton and Christian Haas. "Fairness in Machine Learning: A Survey". In: *arXiv:2010.04053 [cs, stat]* (Oct. 4, 2020). arXiv: 2010.04053. URL: http://arxiv.org/abs/2010.04053 (visited on 06/20/2021) (cit. on p. 9).

[Cho *et al.* 2020] Jaeho Cho, Saifuddin Ahmed, Martin Hilbert, Billy Liu, and Jonathan Luu. "Do Search Algorithms Endanger Democracy? An Experimental Investigation of Algorithm Effects on Political Polarization". In: *Journal of Broadcasting & Electronic Media* 64.2 (May 1, 2020). Publisher: Routledge _eprint: https://doi.org/10.1080/08838151.2020.1757365, pp. 150–172. ISSN: 0883-8151. DOI: 10.1080/08838151.2020.1757365. URL: https://doi.org/10.1080/08838151.2020.1757365 (visited on 12/02/2020) (cit. on p. 8).

[Covington *et al.* 2016] Paul Covington, Jay Adams, and Emre Sargin. "Deep Neural Networks for YouTube Recommendations". In: *Proceedings of the 10th ACM Conference on Recommender Systems*. RecSys '16. New York, NY, USA: Association for Computing Machinery, Sept. 7, 2016, pp. 191–198. ISBN: 978-1-4503-4035-9. DOI: 10.1145/2959100.2959190. URL: https://doi.org/10.1145/2959100.2959190 (visited on 11/08/2020) (cit. on p. 7).

[Dash *et al.* 2019] Abhisek Dash, Animesh Mukherjee, and Saptarshi Ghosh. "A Network-centric Framework for Auditing Recommendation Systems". In: *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*. IEEE INFOCOM 2019 - IEEE Conference on Computer Communications. ISSN: 2641-9874. Apr. 2019, pp. 1990–1998. DOI: 10.1109/INFOCOM.2019.8737486 (cit. on p. 9).

[Faddoul *et al.* 2020] Marc Faddoul, Guillaume Chaslot, and Hany Farid. "A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos". In: *arXiv:2003.03318 [cs]* (Mar. 6, 2020). arXiv: 2003.03318. URL: http://arxiv.org/abs/2003.03318 (visited on 12/02/2020) (cit. on p. 8).

[Giller 2012] Graham L. Giller. *The Statistical Properties of Random Bitstreams and the Sampling Distribution of Cosine Similarity*. SSRN Scholarly Paper ID 2167044. Rochester, NY: Social Science Research Network, Oct. 25, 2012. DOI: 10.2139/ssrn.2167044. URL: https://papers.ssrn.com/abstract=2167044 (visited on 10/29/2020) (cit. on p. 9).

[Goldberg *et al.* 1992] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. "Using collaborative filtering to weave an information tapestry". In: *Communications of the ACM* 35.12 (Dec. 1, 1992), pp. 61–70. ISSN: 0001-0782. DOI: 10.1145/138859.138867. URL: https://doi.org/10.1145/138859.138867 (visited on 02/28/2021) (cit. on p. 2).

[Guy *et al.* 2010] Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, and Erel Uziel. "Social media recommendation based on people and tags". In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '10. New York, NY, USA: Association for Computing Machinery, July 19, 2010, pp. 194–201. ISBN: 978-1-4503-0153-4. DOI: 10.1145/1835449.1835484. URL: https://doi.org/10.1145/1835449.1835484 (visited on 10/29/2020) (cit. on p. 7).

[Harper and Konstan 2015] F. Maxwell Harper and Joseph A. Konstan. "The MovieLens Datasets: History and Context". In: *ACM Transactions on Interactive Intelligent Systems* 5.4 (Dec. 22, 2015), 19:1–19:19. ISSN: 2160-6455. DOI: 10.1145/2827872. URL: https://doi.org/10.1145/2827872 (visited on 02/19/2021) (cit. on p. 14).

[He *et al.* 2016] Chen He, Denis Parra, and Katrien Verbert. "Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities". In: *Expert Systems with Applications* 56 (Sept. 1, 2016), pp. 9–27. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2016.02.013. URL: http://www.sciencedirect.com/science/article/pii/S0957417416300367 (visited on 10/29/2020) (cit. on p. 7).

[Hosseinmardi *et al.* 2020] Homa Hosseinmardi *et al.* "Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube". In: *arXiv:2011.12843 [cs]* (Nov. 25, 2020). arXiv: 2011.12843. URL: http://arxiv.org/abs/2011.12843 (visited on 11/30/2020) (cit. on p. 9).

[Huszár *et al.* 2021] Ferenc Huszár *et al.* "Algorithmic Amplification of Politics on Twitter". In: *arXiv:2110.11010 [cs]* (Oct. 21, 2021). arXiv: 2110.11010. URL: http://arxiv.org/abs/2110.11010 (visited on 11/09/2021) (cit. on p. 10).

[Jiang *et al.* 2019] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. "Degenerate feedback loops in recommender systems". In: *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019), pp. 383–390. ISSN: undefined. DOI: 10.1145/3306618.3314288. URL: https://www.mendeley.com/catalogue/c07bc3f3-281c-3ccb-b9b9-48a7fc7f2ece/?articleTrace=AAABsG80nAYfTD4HP10XPYWudyHkNafVnqCvn6Ig2oJqCCpBaS_wu6VTxyfAdMsg63hV_i76srILGk_K-cUnjMzRq5sM_flieDZo9zBeJAIulo-tCDBDZiQYD3Jhct1NDiXtVWxkivPLBiGtHYoWzRNabD2EjXxPO8EqBV5pBut6uzx23J9_1_QRQpa3fNNitIQHj3gF94ExSAlm6lVuHblzmuV-qpqRkgYPJpDNO-YfrtntvwAP3FRvgIsLQ8DL7-XY9fuXMxGeGIiTOkcbf3rc51ANKDonO9IKWvAJrUCoz6nKynOZYymZ1kv02PQGGXqFSeD_H4a4JkpyYSFqc48QLduCREyiXbme1KGk-_QCXMPu6y5-tfW_h-W28T_287qzhCx7UPZTvDK_nqvzK1rGt0_cZUz2BU34lBGHj6PZWG94rHR75PTtTSfEeDyagem8BxRApr-9q05eUHsdan8x_JPgxJjvEzTchxv958r6-dElNxQUNyBDwdUgxLQvLgUrfSPK8DhlqxWL7T2s4nC560IYd4Nt2toQ4-FoHbrFrfejqDZhGlLO8HX15Ti9Ue6-kp&dgcid=raven_md_suggest_email (visited on 04/20/2021) (cit. on p. 10).

[Kunaver and Požrl 2017] Matevž Kunaver and Tomaž Požrl. "Diversity in recommender systems - A survey". In: *Knowledge-Based Systems* 123 (May 1, 2017), pp. 154–162. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2017.02.009. URL: http://www.sciencedirect.com/science/article/pii/S0950705117300680 (visited on 10/29/2020) (cit. on p. 7).

[Lecher and Yin 2022] Colin Lecher and Leon Yin. *One Year After the Capitol Riot, Americans Still See Two Very Different Facebooks - The Markup*. Section: Citizen Browser. URL: https://themarkup.org/citizen-browser/2022/01/06/one-year-after-the-capitol-riot-americans-still-see-two-very-different-facebooks (visited on 01/24/2022) (cit. on p. 11).

[Ledwich and Zaitsev 2019] Mark Ledwich and Anna Zaitsev. "Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization". In: *arXiv:1912.11211 [cs]* (Dec. 24, 2019). arXiv: 1912.11211. URL: http://arxiv.org/abs/1912.11211 (visited on 11/03/2020) (cit. on p. 10).

[Mansoury *et al.* 2020] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. "Feedback Loop and Bias Amplification in Recommender Systems". In: *arXiv:2007.13019 [cs]* (July 25, 2020). arXiv: 2007.13019. URL: http://arxiv.org/abs/2007.13019 (visited on 06/20/2021) (cit. on p. 10).

[Matakos *et al.* 2020] A. Matakos, C. Aslay, E. Galbrun, and A. Gionis. "Maximizing the Diversity of Exposure in a Social Network". In: *IEEE Transactions on Knowledge and Data Engineering* (2020). Conference Name: IEEE Transactions on Knowledge and Data Engineering, pp. 1–1. ISSN: 1558-2191. DOI: 10.1109/TKDE.2020.3038711 (cit. on p. 8).

[Möller *et al.* 2018] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. "Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity". In: *Information, Communication & Society* 21.7 (July 3, 2018). Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2018.1444076, pp. 959–977. ISSN: 1369-118X. DOI: 10.1080/1369118X.2018.1444076. URL: https://doi.org/10.1080/1369118X.2018.1444076 (visited on 02/05/2021) (cit. on p. 10).

[Munger and Phillips 2020] Kevin Munger and Joseph Phillips. "Right-Wing YouTube: A Supply and Demand Perspective". In: *The International Journal of Press/Politics* (Oct. 21, 2020). Publisher: SAGE Publications Inc, p. 1940161220964767. ISSN: 1940-1612. DOI: 10.1177/1940161220964767. URL: https://doi.org/10.1177/1940161220964767 (visited on 12/02/2020) (cit. on p. 10).

[M. H. Ribeiro *et al.* 2020] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. "Auditing radicalization pathways on YouTube". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. New York, NY, USA: Association for Computing Machinery, Jan. 27, 2020, pp. 131–141. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372879. URL: https://doi.org/10.1145/3351095.3372879 (visited on 10/29/2020) (cit. on p. 10).

[P. V. Ribeiro 2021] Paulo Victor Ribeiro. *Como a extrema direita burla punições do YouTube - e o Google finge que não vê*. The Intercept. Apr. 19, 2021. URL: https://theintercept.com/2021/04/19/como-a-extrema-direita-burla-punicoes-do-youtube-e-o-google-finge-que-nao-ve/ (visited on 08/09/2021) (cit. on p. 11).

[Ricci *et al.* 2011] Francesco Ricci, Lior Rokach, and Bracha Shapira. "Introduction to Recommender Systems Handbook". In: *Recommender Systems Handbook*. Ed. by Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. Boston, MA: Springer US, 2011, pp. 1–35. ISBN: 978-0-387-85820-3. DOI: 10.1007/978-0-387-85820-3_1. URL: https://doi.org/10.1007/978-0-387-85820-3_1 (visited on 02/28/2021) (cit. on pp. 3, 16).

[Roth *et al.* 2020] Camille Roth, Antoine Mazières, and Telmo Menezes. "Tubes and bubbles topological confinement of YouTube recommendations". In: *PLOS ONE* 15.4 (Apr. 21, 2020). Publisher: Public Library of Science, e0231703. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0231703. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231703 (visited on 01/11/2021) (cit. on p. 9).

[Sarwar *et al.* 2001] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. "Item-based collaborative filtering recommendation algorithms". In: *Proceedings of the 10th international conference on World Wide Web*. WWW '01. New York, NY, USA: Association for Computing Machinery, Apr. 1, 2001, pp. 285–295. ISBN: 978-1-58113-348-6. DOI: 10.1145/371920.372071. URL: https://doi.org/10.1145/371920.372071 (visited on 02/28/2021) (cit. on p. 15).

[Séraphin *et al.* 2017] Alava Séraphin, Frau-Meigs Divina, and Hassan Ghayda. *Youth and violent extremism on social media: mapping the research*. Google-Books-ID: PTRCDwAAQBAJ. UNESCO Publishing, Dec. 4, 2017. 167 pp. ISBN: 978-92-3-100245-8 (cit. on p. 4).

[Sinha *et al.* 2017] Ayan Sinha, David F. Gleich, and Karthik Ramani. "Deconvolving Feedback Loops in Recommender Systems". In: *arXiv:1703.01049 [cs]* (Mar. 3, 2017). arXiv: 1703.01049. URL: http://arxiv.org/abs/1703.01049 (visited on 11/08/2020) (cit. on p. 10).

[Sîrbu *et al.* 2019] Alina Sîrbu, Dino Pedreschi, Fosca Giannotti, and János Kertész. "Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model". In: *PLOS ONE* 14.3 (Mar. 5, 2019). Publisher: Public Library of Science, e0213246. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0213246. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0213246 (visited on 10/29/2020) (cit. on p. 9).

[Stoica 2020] Ana-Andreea Stoica. "Algorithmic Fairness for Networked Algorithms". In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '20. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, May 5, 2020, pp. 2214–2216. ISBN: 978-1-4503-7518-4. (Visited on 10/29/2020) (cit. on p. 8).

[Stoica and Chaintreau 2019] Ana-Andreea Stoica and Augustin Chaintreau. "Hegemony in Social Media and the effect of recommendations". In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW '19. New York, NY, USA: Association for Computing Machinery, May 13, 2019, pp. 575–580. ISBN: 978-1-4503-6675-5. DOI: 10.1145/3308560.3317589. URL: https://doi.org/10.1145/3308560.3317589 (visited on 10/29/2020) (cit. on p. 8).

[Stoica, Riederer, *et al.* 2018] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. "Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity". In: *Proceedings of the 2018 World Wide Web Conference*. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, Apr. 23, 2018, pp. 923–932. ISBN: 978-1-4503-5639-8. DOI: 10.1145/3178876.3186140. URL: https://doi.org/10.1145/3178876.3186140 (visited on 10/29/2020) (cit. on pp. 8, 14).

[Su *et al.* 2016] Jessica Su, Aneesh Sharma, and Sharad Goel. "The Effect of Recommendations on Network Structure". In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, Apr. 11, 2016, pp. 1157–1167. ISBN: 978-1-4503-4143-1. DOI: 10.1145/2872427.2883040. URL: https://doi.org/10.1145/2872427.2883040 (visited on 10/29/2020) (cit. on p. 9).

[Tangherlini *et al.* 2020] Timothy R. Tangherlini, Shadi Shahsavari, Behnam Shahbazi, Ehsan Ebrahimzadeh, and Vwani Roychowdhury. "An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web". In: *PLOS ONE* 15.6 (June 16, 2020). Publisher: Public Library of Science, e0233879. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0233879. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0233879 (visited on 10/29/2020) (cit. on p. 8).

[*TensorFlow Recommenders* 2021] *TensorFlow Recommenders*. TensorFlow. URL: https://www.tensorflow.org/recommenders (visited on 03/01/2021) (cit. on p. 19).

[Wong 2021] Julia Carrie Wong. "How Facebook let fake engagement distort global politics: a whistleblower's account". In: *The Guardian* (Apr. 12, 2021). ISSN: 0261-3077. URL: https://www.theguardian.com/technology/2021/apr/12/facebook-fake-engagement-whistleblower-sophie-zhang (visited on 08/09/2021) (cit. on p. 11).

[Yao *et al.* 2021] Sirui Yao *et al.* "Measuring Recommender System Effects with Simulated Users". In: *arXiv:2101.04526 [cs]* (Jan. 12, 2021). arXiv: 2101.04526. URL: http://arxiv.org/abs/2101.04526 (visited on 06/20/2021) (cit. on p. 9).

[Zhao *et al.* 2019] Zhe Zhao *et al.* "Recommending what video to watch next: a multitask ranking system". In: *Proceedings of the 13th ACM Conference on Recommender Systems*. RecSys '19. New York, NY, USA: Association for Computing Machinery, Sept. 10, 2019, pp. 43–51. ISBN: 978-1-4503-6243-6. DOI: 10.1145/3298689.3346997. URL: https://doi.org/10.1145/3298689.3346997 (visited on 11/05/2020) (cit. on pp. 3, 7).

[Ziegler 2004] Cai-Nicolas Ziegler. *Book-Crossing Dataset*. Sept. 2004. URL: http://www2.informatik.uni-freiburg.de/~cziegler/BX/ (visited on 03/01/2021) (cit. on p. 14).