Inescapable bias

The role of recommender systems in social media radicalization

Caio Lente

Report presented to the
Institute of Mathematics and Statistics
of the University of São Paulo
for the Master of Science
qualifying examination

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

São Paulo
February 10th, 2021

Inescapable bias

The role of recommender systems in social media radicalization

Caio Lente

This is the original version of the qualifying text 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***. Exame de Qualificação (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. **Inescapable bias:** ***The role of recommender systems in social media radicalization***. Qualifying Exam (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 3

1.3 Radicalization 4

1.4 Hypothesis 5

2 Literature review **7**

3 Proposal and Preliminary Experiments **11**

3.1 Experiments 12

3.2 Further Experiments and Schedule 16

Appendices

Annexes

References **17**

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. The rest of Chapter 1 is dedicated to core concepts covered in the rest of the work, ending in subsection 1.4, which tackles the main hypothesis of this dissertation. Chapter 2 contains the literature review, Chapter 3 explains the experiments already conducted and their results, and Chapter 4 is about next steps.

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 11 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.

Some social media services, very much in agreement with McLuhan's teachings, have what could be considered a "style". Instagram's content, for example, tends toward more personal (i.e. egoic) photos and videos. As of November 28th, 2020, six of the top 20 most-liked posts on Instagram are from american socialite Kylie Jenner, consisting of four photos of her daughter and two of her ex-boyfriend. Even though there are many different niches inside Instagram, personal posts seam to have an edge over other kinds of content.

Twitter, unlike most other social networks, allows for asymmetrical connections, meaning users can follow profiles without being followed back. This enables the emergence of Twitter communities (e.g. Fintwit, Black Twitter) that can be largely self referential and/or organized around certain subjects. Facebook users, on the other hand, can belong to groups, user-moderated profiles that might revolve around any particular topics of interest; there are groups that organize pet owners and groups that organize neonazis.

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 recomender algorithms) first appeared in 1992 under the name "collaborative filtering", even though that term nowadays refers to a subclass of recommender systems. 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 rcommended items based on the ratings the algorithms estimated the user would give to those items.

Most recommender systems follow 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. One such algorithm might use the genres of watched movies in order to recommend new ones, while another might analyse 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. Hybrid filtering usually mix collaborative methods with either content-based or demographic filtering.

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. 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.

1.3 Radicalization

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:

- 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. This second point is of note: content that appears urgent and leaves the user fearful (for their live, their community, or their identity) is more engaging and, therefore, 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.

1.4 Hypothesis

As explained in the previous sections, social networks' recommendation algorithms might play a significant role in radicalizing users. This could, at least in part, explain the recent surge in popularity that far-right ideologies have enjoyed over the last few years. If true, this is an existential threat to modern democracies that should be addressed as soon as possible.

This dissertation aims to explore the radicalization pipeline hypothesis and, more specifically, understand the mechanisms through which recommender systems can end up suggesting extreme content to regular users. 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 they might lead to "fixed points" in an algorithm's phase space.

In short, the main goal is to test the pipeline hypothesis in a setting where recommendation algorithms are modeled as dynamical systems. This will allow for a better understanding of how these systems behave in the wild, possibly taking the user in a radicalizing "trip" through the space of all possible items.

Chapter 2

Literature review

There are three types of work that are relevant to the current topic: general literature about recommender systems, evidence for the radicalization hypothesis, and evidence against the radicalization hypothesis. 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). Original research is, therefore, still required before a final verdict is issued.

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.

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." The results presented in this paper are in line with the recommender system radicalization hypothesis. 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 Stoica and Chaintreau (2019), Matakos *et al.* (2020) also propose a novel recommendation 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 thees 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.

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.

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 seminal article in the field, by 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 doesn't prove that the pipeline exists, but is a strong argument for its existence.

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.

Chapter 3

Proposal and Preliminary Experiments

As discussed in the previous chapters, understanding how social networks recommend content to users is central to the debate around the recent waves of political polarization and radicalization that have been taking over many developing and developed countries alike. 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 work. This means that studying a social network's recommendation technique has inherent limitations.

Since the life and blood of almost all social media platforms revolve around their recommendations, most of the algorithms currently employed by these 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 (which is more people than any country in the planet), 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. YouTube claims only a small fraction of its content is political in nature, but that doesn't mean it is not enough to spread across the internet and help radicalize users the world over. 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 extreme content, most of the non-English-speaking world would still not be impacted by their policy changes.

Even with a quickly growing body of research, further studies are desperately 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 the last 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.

This leads right into the goal of the present report. The dissertation to be presented as a result of this program aims 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 recommendation algorithms always create "filter bubbles", suggesting ever more engaging videos about a certain topic of interest to a user, and possibly sending them on a radicalization spiral if that topic is related to politics or other contentious subjects.

3.1 Experiments

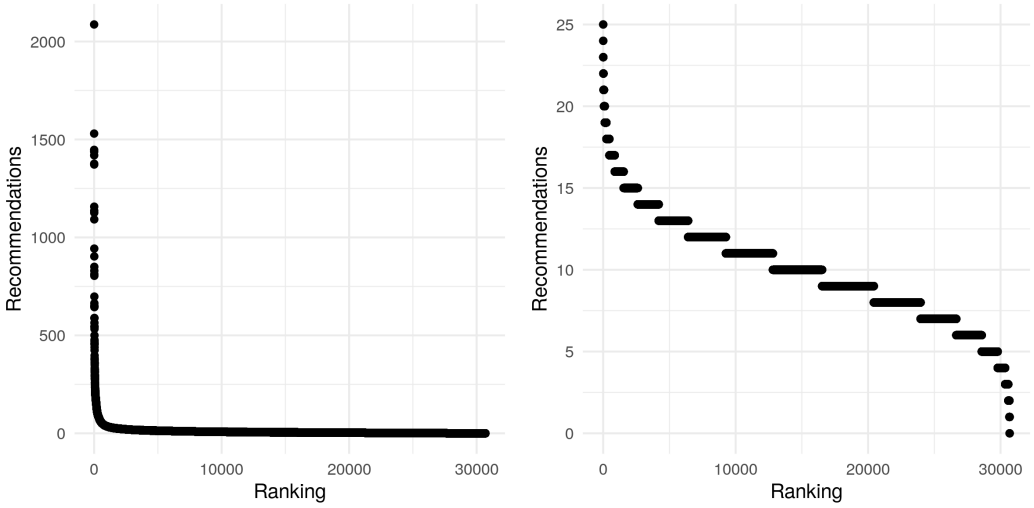
Some preliminary experiments have already been conducted in order to gather some evidence in favor or against the main hypothesis being tested. If these experiments had failed, then there would be no reason to continue pursuing this argumentative path. In total, fifteen different recommendation models were trained and analyzed, with each visualization below representing one of these models.

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 would get transferred on to the model; the objective here is understanding the algorithm by itself without external influences.

The chosen type of recommender system to be trained was, therefore, a 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.

The main recommendation model (henceforth referred to as the "vanilla" model) was trained with the MovieLens dataset (Harper and Konstan, 2015). 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 k most similar vectors.

Once the model was ready, the analysis started: the algorithm was asked to generate a list of the top $k = 10$ most similar movies to each movie; since there were 30689 movies in the dataset, this created a list of 30689×10 movie IDs. After this computationally intensive calculation, the occurrence of each ID was counted and ranked accordingly, namely, the movie ranked number 1 was to be the movie featured the most times in the set of all recommendations, and so forth for every other rank. The result of this process can be seen in Figure 3.1a.

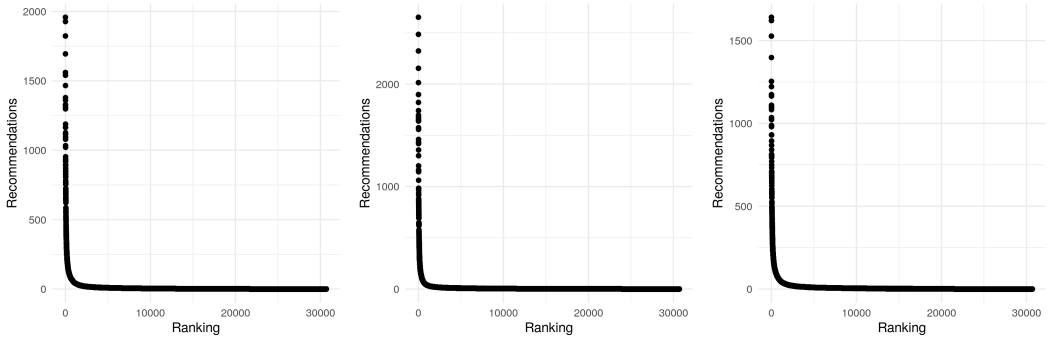


(a) *Rec. based on movie metadata.* **(b)** *Random recommendations.*

Figure 3.1: *Vanilla recommendation model and control.*

For comparison, Figure 3.1b is the same visualization for the trivial model (sampling k movies at random when asked for a recommendation). The number of times each movie appeared in the final list of all recommendations averaged, evidently, k, 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. However, the distribution of the vanilla model differs immensely: the movie ranked number 1 appeared more than 2000 times in the final list, with an almost exponential decrease in the number of appearances from then on.

In order to better understand this phenomenon, more models were trained and analyzed. Since all visualizations present exactly the same quantities, obtained in the exact same way, all of them can be compared.

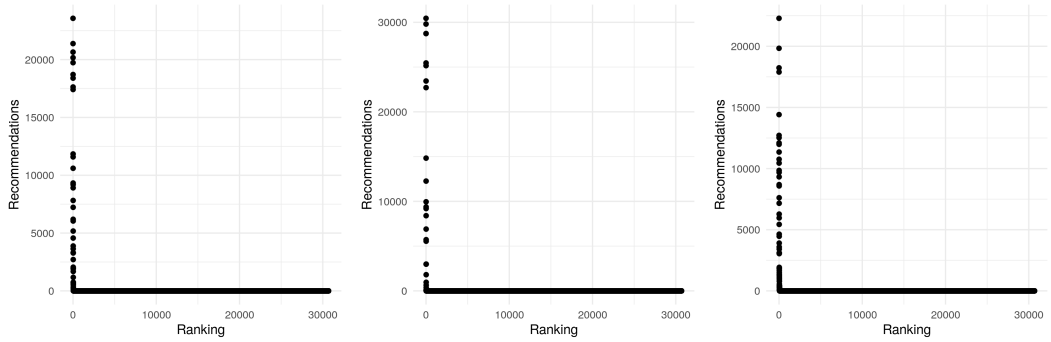


(a) *Cutoff* $n \leqslant 5$. **(b)** *Cutoff* $n \leqslant 2$. **(c)** *Cutoff* $n \leqslant 8$.

Figure 3.2: *Vanilla model with cutoff point.*

The first variation on the vanilla model that was experimented on was creating cutoff points for words to be included in the vector representation of the movies. Since most words appear only once (given that many are director and cast names), the few movies whose metadata contained popular terms could have been favored by the recommendation

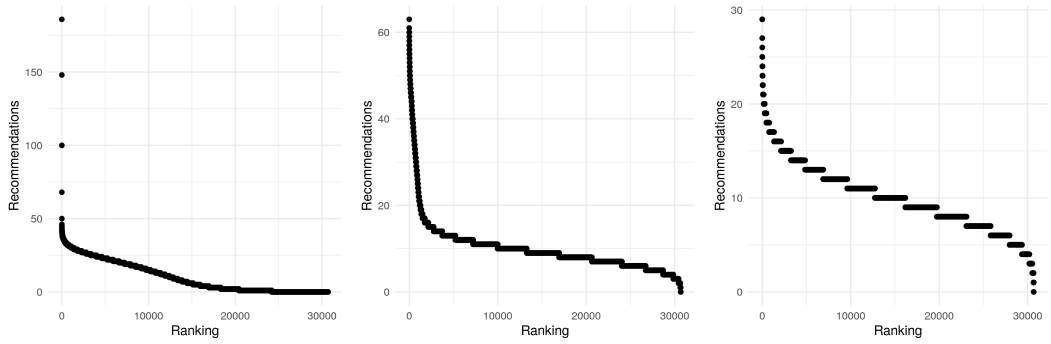
algorithm. Three cutoff points were created, and the model was retrained with datasets that only contained words that appeared at least twice, five times, and eight times. The results can be seen in Figure 3.2 and, aside from variations in the y-intercept, all plots are qualitatively very similar to Figure 3.1a.



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

Figure 3.3: *Using distances instead of cosine similarity.*

The second validation experiment involved using distance metrics instead of cosine similarity. 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 that the rest. As attested by Figure 3.3, this was not the case. No other similarity metrics were used because cosine seems to be the most popular one used in simple recommender systems. More experiments have still to be conducted on this front, specially comparing what IDs belong to the group of top-recommended movies for each metric and if there is intersection between them.

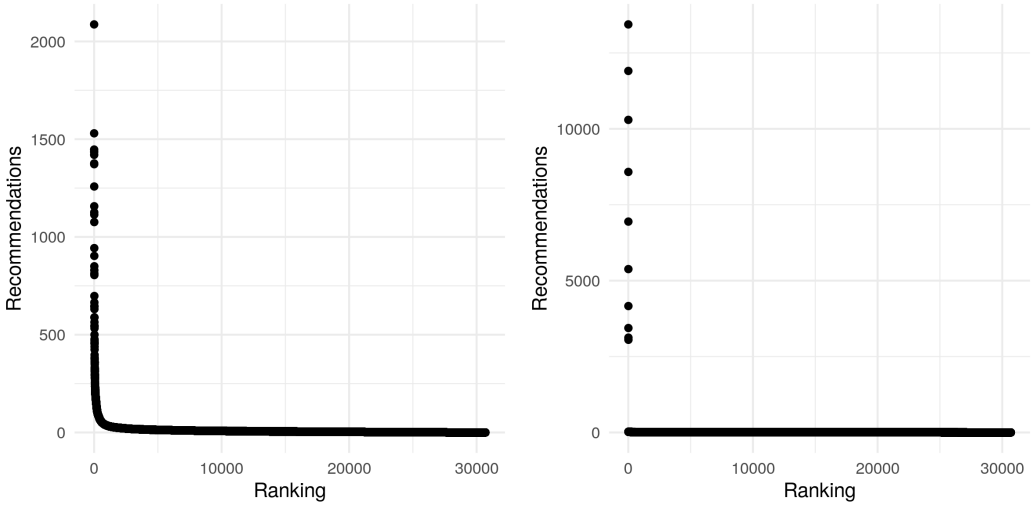


(a) $P(w) = P(V_w)$. **(b)** $P(w) = 10 \times P(V_w)$. **(c)** $P(w) = 100 \times P(V_w)$.

Figure 3.4: *Random sample of words from metadata.*

At this point it is safe to say that the exponential decay in recommendation frequencies is not spurious and must have a clear cause. A hypothesis that was later mostly confirmed involved 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. Figure 3.4 displays the curves of three different models trained from random data; for each, the propability of an element being non-zero in the vector representation of a movie was the average probability that an arbitrary element

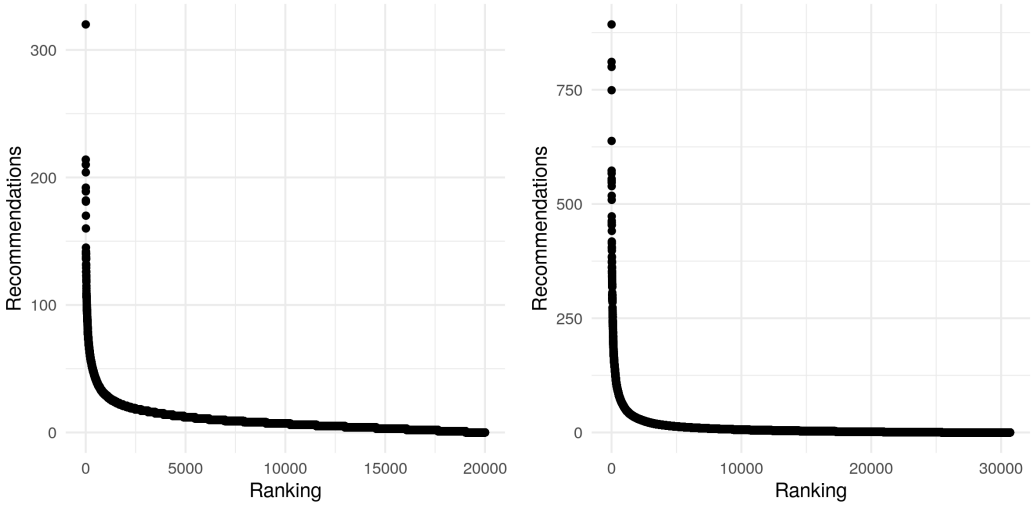
of the MovieLens dataset was non-zero (times 1, 10, and 100). This was equivalent to creating random metadata for the movies where the probability of a word occurring was approximately 1.54×10^{-4}, 1.54×10^{-3}, and 1.54×10^{-2} respectively. The results support the aforementioned hypothesis.



(a) *Artificial movie.* **(b)** *Long and narrow.*

Figure 3.5: *Sanity checks.*

Figure 3.5 showcases two sanity checks. Figure 3.5a was a model trained with the vanilla dataset, but with the addition of an artificial movie created as a combination of the metadata from other movies favored by the recommendation algorithm. As expected, this movie also featured in the top-recommended subset. Figure 3.5b comes from a model trained on random data generated in a similar fashion to the model in Figure 3.4a, except each vector could only have 15,000 elements instead of 55,681 as with the vanilla model (which is why it is called "long and narrow"). The patter observed before persisted.



(a) *Book dataset.* **(b)** *Simulation of vanilla.*

Figure 3.6: *Confirmation of hypothesis.*

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 exponentially more recommended than the rest as long as the data was sparse. Figure 3.6a represents the same recommendation algorithm applied to another dataset, the Book-Crossing Dataset. Figure 3.6b contains the results of the model applied to another random dataset, 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.

3.2 Further Experiments and Schedule

As previously mentioned, more experiments are still necessary in order to identify exactly what is the nature of the bias detected in the MovieLens study. For example, it would be interesting to perform an analysis of what movies are the most recommended in each case and whether the subset of top-recommended movies is roughly consistent overall. Better understanding of the literature about recommender systems would also be of great help in searching for possible explanations to this phenomenon.

The most important experiment that hasn't yet been conducted regards the dynamic nature of recommendation algorithms. Using Google's newly released TensorFlow Recommenders library it might be possible to gather data about what happens to a system's recommendations as users follow (or don't follow) its suggestions, specially when talking about deep learning models used by social networks like YouTube. This could either support or disprove the hypothesis that these algorithms could be suggesting ever more extreme content in order to engage users, creating a filter bubble effect. Literature about this topic is still scarce, but every day new papers about the effects that social media has on Democracy are being released, so following up on popular publications is of the utmost importance.

These two main tasks, alongside writing the final dissertation, will be the focus of the comming semester. A full schedule can be seen below in Figure 3.7.

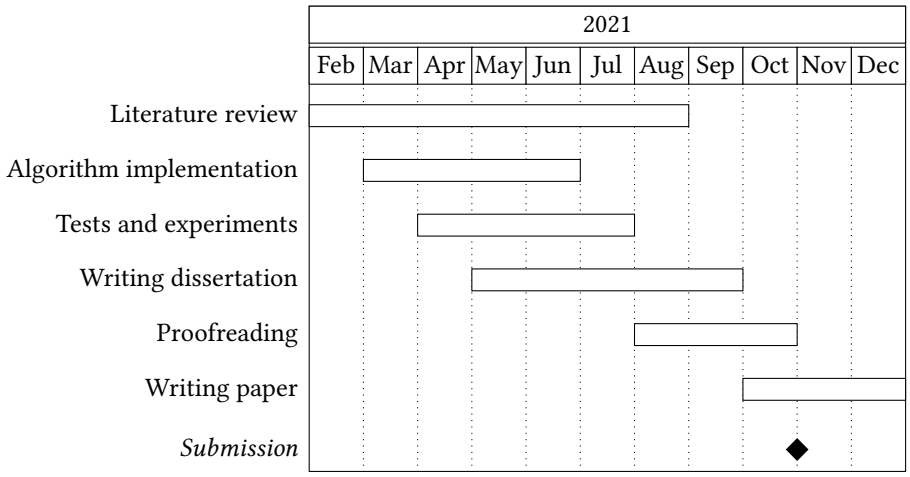


Figure 3.7: *Schedule.*

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).

[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).

[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).

[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).

[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. 12).

[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).

[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).

[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).

[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).

[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. 9).

[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. 9).

[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).

[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, 12).

[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. 7).

[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 p. 7).