2019

Empirical studies of factors affecting opinion dynamics

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

EMPIRICAL STUDIES OF FACTORS AFFECTING OPINION DYNAMICS

by

LARISSA PINHEIRO SPINELLI

B.S., Universidade Federal do Rio Grande do Norte, 2009
M.S., Universidade Federal do Rio de Janeiro, 2011

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2019

Approved by

First Reader

Mark Crovella, PhD
Professor of Computer Science

Second Reader

Evimaria Terzi, PhD
Associate Professor of Computer Science

Third Reader

George Kollios, PhD
Professor of Computer Science

Fourth Reader

Andrei Lapets, PhD
Associate Professor of the Practice of Computer Science

All life is an experiment. The more experiments you make the better.
Ralph Waldo Emerson

Acknowledgments

I want to acknowledge the immense support that I had from family, mentors, friends, and colleagues during my Ph.D. journey.

First, I would like to thank the great support and mentorship of my adviser Mark Crovella that guided with patience along the Ph.D. years. It was a pleasure sharing this journey and learn a lot with his immense sense of curiosity and intelligence. I also would like to thank the CS professors that taught me classes, my thesis committee members by the discussions and suggestions on my thesis and, all the CS staff members that made my life easier. Also, I would like to thank Daniel Ratton, my Master's adviser, who pushed and helped me to start this (international) Ph.D. journey.

I'm eternally grateful to my family (core and extended) which love and, the support gave me the strength to pursue my dreams and persevere into this hard journey. I would like specially dedicate this work to my parents, Regina and Geraldo, that always believe in me and taught me to love the knowledge journey. Also, my siblings, Luciana, Luisa, and Daniel, that inspired me to be brave and see life plurality though our so different paths. Plus, my "American" family, Heloisa e Roberta by the close support.

I'm glad and thankful to have shared the course of my Ph.D. with some fantastic colleagues between classes, lunches, coffee/teas, chocolates, drinks, and some many other grad student moments. It impossible list them all, but I will risk thank at least some by name: Giovanni by the many discussions and conversations about classes, science, and life; Mike that not only shared classes and computer discussions but some adventures of hikes and climbings; Maryam, my lunch buddy of many years; Natali, by the lovely presence; and some many other such as Matt, Harry, Nabeel, Sarah, Sofia, Hanna, Behzad, Sanaz, Mehrnoosh, Athina, Yida, Alex, Ye, Ying, Flavio and, Yuefang. I could not also forget my dear officemates: Gonca, Baichuan, and Bashir.

Last but not least, I would like to thank amazing friends that help me in my journey outside the CS department. To my many flatmates to their friendship and care: Magda, Josiane, Leisa, Ana Lucia, Cansu, Tibi, Carol, Leyla, Fulya, Daniella, Alice, Luke, Mike, and, Zaki. My dear friends and accompanies of sports or adventures: Ida, Nag, Ida, Nag, Andrew, Carsten, Javad, Nestor, Carlos and, Emir. And some dear Brazilians that helped to cope with my homesick: Flavia, Rodrigo, Lucy, Monica, Victor, Mary, Jacques, Daniel, Bruna and, Diogo. Moreover, I'm also thankful to many dear friends that from far supported me.

EMPIRICAL STUDIES OF FACTORS AFFECTING OPINION DYNAMICS

LARISSA PINHEIRO SPINELLI

Boston University, Graduate School of Arts and Sciences, 2019

Major Professor: Mark Crovella, PhD
Professor of Computer Science

ABSTRACT

The advent of new online services has an enormous potential to impact the opinion of users. Two main drivers of this impact are crowdsourced evaluations and ratings, and algorithmically-chosen recommendations. However, understanding the relationship between these systems and their impacts is very challenging due the complex nature of recommender systems and due to the heterogeneous nature of crowdsourced reviews. In this thesis, we explore how these two drivers affect opinion dynamics with respect to two potential impacts: reliability of information and polarization of user opinion. First, we analyze the reliability of online ratings. By performing an empirical analysis of a large corpus of online ratings, we point out how different influences such as shifts in population or platform characteristics are correlated with changes in the perception of an item over time. Second, we investigate polarization in the context of recommender systems. We define three metrics – intensity, simplification, and divergence – to capture essential traits of user opinions and explore how they vary in a closed-loop with recommender systems. Finally, we examine reliability in recommendations via an empirical exploration on YouTube. We quantify changes in the nature of the recommended content, and we show how YouTube recommendations lead users – especially privacy-seeking users – away from reliable information. Taken together, these studies shed light on important factors that affect how user opinion is shaped by online systems.

Contents

List of Tables

List of Figures

List of Abbreviations

ACM	Association for Computing Machinery
AI	Artificial Intelligence
AZ	Amazon
BX	Book-Crossing
CS	Computer Science
CrowdRec	Crowdsourcing and Human Computation for RecSys
DiveRS	Novelty and Diversity in Recommender System
EPL	Europhysics Letters
IEEE	Institute of Electrical and Electronics Engineers
ICASSP	Conference on Acoustics, Speech, and Signal Processing
ICF	Item-based Collaborative Filtering
IMC	Internet Measurement Conference
ML	Movie Lens
MT	Movie Tweetings
PLoS	Public Library of Science
PPR	Personalized PageRank
RAN	Random
Rec	Recommendation
RecSys	Recommender Systems
RT	Rotten Tomatoes
SIGIR	Research and Development in Information Retrieval
SIGKDD	Knowledge Discovery and Data Mining
SIGMOD	Knowledge Discovery and Data Mining
SVD	Singular Value Decomposition
TOIS	Transactions on Information Systems
TopiCS	Topics in Cognitive Science
WOM	Word-of-Mouth
WSDM	Conference on Web Search and Data Mining
WWW	World Wide Web Conference
YT	YouTube

Chapter 1

Introduction

The Internet has revolutionized information access and the structure of media. The removal of physical barriers has made it possible to access tremendous amounts of information and products. This abundance of choices presents issues. Users must decide which goods to purchase, which news to access, which movies to watch, and which products to consume. In this context, several online services utilize recommendation systems or crowdsourced-reviews as a way to assist users to sort out this overwhelming set of choices.

As a result, crowdsourced reviews have become a significant information source for consumers, retailers, manufacturers, and producers. Decisions such as the next place to eat, an enjoyable movie to watch, a vacation destination, or items/services that are worth to purchasing are often made after users access and compare online reviews to estimate perceived quality and reduce uncertainty about goods. Because online reviews have the potential to support decision making, it is essential to consider the extent to which platforms, users, or the heterogeneity of goods evaluated could add or amplify biases in information coming from these reviews.

In addition, recommender systems are a pervasive part of the online experience. They curate, for instance, the results of search queries, the selection of news, the music that will be played, the friends' updates that will show up on social media, the next video that will be watched, or a user's next purchase. These algorithmically-chosen recommendations assist in navigating on the torrent of choices offered by the online services. However, the design of recomendation algorithms – which goals they will optimize – can impact their recommendations immensely and can amplify biases.

In essence, every technology has an interface, and when this interface is shaping the information accessed, it has the potential to impact users' perceptions of the world. Therefore, the general goal of this dissertation is to enrich the understanding of how two main drivers of impact – crowdsourced reviews and algorithmically-chosen recommendations – affect users' opinions. With respect to impact, we explore two dimensions. The first dimension relates to the value of the recommended content, i.e., content reliability, the second dimension is associated with the spectrum of content made available, i.e., content polarization.

Hence we are concerned with two drivers of impact (crowdsourced reviews and algorithmically-chosen recommendations) and two dimensions of impact (reliability and polarization). Table 1.1 summarizes the organization of this dissertation concerning these combinations. The rest of this chapter will provide an overview of each of these topics, and Chapter 2 will review related works. Then, the following chapters we will address each of the contributions shown in Table 1.1 in turn.

Table 1.1: Drivers of impact and Dimensions of Impact

Dimensions	*Drivers of Impact*	
	Reviews	**Recommendations**
Reliability	Chapter 3	Chapter 5
Polarization		Chapter 4

1.1 Crowdsourced Reviews

The first part of the dissertation, Chapter 3, analyzes factors that impact ratings seen in crowdsourced reviews.

1.1.1 Reliability

In the context of online ratings, the reliability dimension translates to the initial question of how to interpret a rating. Additionally, noting that there is generally temporal variation in these scores, reliability concerns the question of why ratings change. These questions do not have simple answers because many forces play a role in the variability of ratings. Ratings of different types of products, for instance, can have distinct variation over time, affecting the way that their value should be assessed. Even for a single type of product, users should consider that the design or services provided by a platform could guide review practices creating specific trends or inserting biases. Moreover, different populations could appraise the quality of a product in different manners.

Our approach to this problem is presented in Chapter 3. We break down ratings into interpretable components and compare differences, either among these components over time, or between carefully selected data. By creating a holistic picture of the forces that combine to determine online rating dynamics, we quantify changes, understand the role played by diverse sources, and in this way shed light on how reliable a rating is.

We develop this approach and apply it to a set of selected datasets. First, we characterize rating dynamics as a whole. By identifying trends, we observe how ratings vary on different platforms. The natural question that follows from there concerns the consistency of platform-specific behavior. We demonstrate the reliability of platform trends by contrasting the same set of items across platforms and by comparing behavior among different categories on the same platform.

Second, using our multi-factor model, we decompose ratings to identify the distinct impact of users, items, and closed-loop interactions. Our results reveal interesting drivers of changes in online ratings. We show that shifts in the population (kind of users) and changes in the perceived quality of the items over time are significant contributors to the ratings' changes. In particular, in some settings, we show that early reviews are mostly made by critics and have a lower score than later reviews, made mainly by the general population. Furthermore, we show how characteristics of the platform affect rating dynamics. One of our results points out that if the platform does not display identified reviews, the perceived quality of items over time tends to decrease. Also, we expose inflation in review scores that support e-commerce and show that in the presence of a recommender system, higher interaction (closed-loop) score items get more reviews.

1.2 Recommendation Systems

In the second part of the dissertation (Chapter 4 and Chapter 5) we analyze how recommender systems impact either the spectrum of information made available (polarization), or content reliability.

1.2.1 Polarization

Another way that users opinion can be impacted is by the diversity or the polarization of content made available to users.

In Chapter 4, we investigate the impact of recommender systems on opinion polarization. Specifically, we investigate how recommender systems can affect the polarity of users' opinions and the nature of this effect. This question becomes more complex when considering that recommender systems are often adaptive, that is, they adjust themselves over time when interacting with users. Accordingly, we investigate then how users' opinions evolve in the dynamical system formed of a recommender system and users in a closed-loop.

As a first step, we define a framework of analysis in which we abstract the system as a matrix of users and items in which values represent the opinion or connection of the user about or with the items. This matrix is changing over time, and we take snapshots in time to evaluate it. Our goal is to evaluate the way that the recommender system impacts users' opinions. That is, do recommender systems reinforce user opinions (make them stronger)? Do they simplify user opinions, making them less diverse? And do they simplify the whole set of user opinions, making them more similar? We assess these questions by defining three metrics – intensity, simplification, and divergence – computed over the users' opinion vectors abstracted from the matrix snapshot.

Our model to analyze polarization in link-based systems is inspired by the model used in a previous theoretical analysis made by (Dandekar et al., 2013). However, we approach the polarization analysis using simulations. We consider three algorithms with different level or randomness (simpleSALSA, simpleICF, and simplePPR) and two users responses (with and without biased assimilation). Using several study cases, we point out the role played by the recommender algorithms, the user response, and the initial conditions on the polarization of opinions. For instance, we show that algorithms that limit the diversity of recommendation while targetting the best recommendation (such as simpleICF and simplePPR) will polarize users' opinions in a broader set of conditions than algorithms less focused on the best recommendation but with higher diversity (like simpleSALSA). Also, the fact that our results contradict previous work reveals how sensitive the system' dynamics are to assumptions and therefore highlights the importance of more extensive analyses like ours.

In the rating-based system, we also use simulations to assess the impact of the start point, recommender algorithm, and user responses. However, given the higher complexity of this system, the initial state is represented there by different datasets. Further, we measure the impact of the recommender system via the contrast of the typical recommender algorithm against a random one. We then present a detailed analysis regarding the impact of the recommender system for each one of our three metrics in time, and for different users' agreement with the system. As a whole, our results attest to how recommender systems polarizes user opinions by their tendency to simplify the user's preferences and to increase the divergence among all users' preferences.

1.2.2 Reliability

In Chapter 5 we explore the impact of recommender systems regarding content reliability. Our goal is to understand and quantify the nature of recommendations focusing on socially-impactful dimensions - notably, the presence of recommendations for reliable versus unreliable information sources.

To approach this problem we conduct an experimental analysis of the nature of the recommended content on YouTube. YouTube is one of the most significant sources of socially-generated information globally, and it is estimated that approximately 70% of its viewing time originates from its recommendations. Consequently, recommendations on YouTube have enormous potential to impact users' opinions worldwide. In recent years, YouTube has been criticized in the popular press for promoting radical, extreme, or unreliable content. In this part of the dissertation, we scrutinize the recommendation behavior of YouTube and contribute a set of tools to analyze its dynamics in time.

Our experimental setup includes the design and implementation of a data collection framework that imitates users who receive recommendations and view videos on YouTube. Our framework also includes the classification of the recommended videos with regard to their reliability.

Using our framework, we demonstrate several unusual behaviors of YouTube recommendations. First, we observe shifts in the types of recommended sources inside a sequence of recommendations, and we show that overall, YouTube recommendations lead users away from trustable sources. Going one step further, we contrast different policies for selecting the next video and show that the "lead away" effect is reinforced by YouTube, that is, YouTube recommendations emphasize (more strongly) less reliable sources.

Furthermore, motivated by the fact that some users value privacy and the protection of their identity, we analyze recommendation sequences for distinct privacy scenarios. Our results reveal a tension between privacy-seeking users and extreme content, we show that users that demand more privacy receive as a counterpart more extreme content.

Next, we investigate the impact of the search query itself, which initiates the sequence of videos recommended. Our results show that although shifts in reliability vary with queries, in most of the topics studied, the Youtube "lead away" effect is present.

Fortunately, during the time of our experiments, YouTube disclosed a change in its recommendation policy to address criticism about unreliable recommendations. Our last analysis contrasts data before and after that change and shows that, although after the disclosure the "lead away" effect is reduced, it is still present.

1.3 Roadmap

The subsequent parts of this thesis are structured as follows:

- Chapter 2 reviews work related to this thesis and contrasts differences with this thesis.

- Chapter 3 presents the first part of the dissertation, where we describe a set of tools and results to elucidate reliability in the context of online ratings.

- Chapter 4 presents the second part of the thesis, where we investigate the dynamics of polarization in recommender systems.

- Chapter 5 comprises the third part of the dissertation. In this chapter we explore the reliability of recommendations through an empirical exploration of the nature of YouTube recommendations.

- Chapter 6 reviews the contributions of this dissertation, examines its entanglements and extensions, and brings closing consideration points.

1.4 Related publications

The work presented in this dissertation is related to two published papers and one paper under submission, which are listed below.

Larissa Spinelli and Mark Crovella. *Unravelling Dynamics of Online Ratings.* SMMA 2018.

Larissa Spinelli and Mark Crovella. *Closed-loop Opinion Formation.* WebSci 2017.

Larissa Spinelli and Mark Crovella. *How YouTube leads privacy-seeking users away from reliable information.* Under Submission

Chapter 2

Related Work

Understanding how online experiences can impact users' opinions is a complex problem that has been explored in part by several works. In this thesis, we focus on understanding how the two primary drivers of opinions – crowdsourced reviews and algorithmically-chosen recommendations – impact users' opinions regarding polarization and reliability. We present our work in turns regarding the combination of drivers and dimensions of impact.

In this chapter, we will present our definitions of polarization and reliability contrasting with definitions and models found on related work. Then, we review work related to these topics, and we position our work and contributions with respect to them.

We review works related to reliability, focusing on two causes of unreliability: temporal dynamics and content value. The former refers to how to assess the value of pieces of information that have temporal dynamics. We address unreliability caused by temporal dynamics on Chapter 3 in the context of online ratings. Works associated with unreliability by temporal dynamics are mostly focused on how to model recommendations or interpret online ratings in time.

Next, we examine work associated with polarization. Following our discussion in Chapter 4, we focus on work related to the dynamics of polarization in recommender systems, i.e., work associated with how the process of opinion formation leads to polarization.

The latter cause of unreliability (content value) refers to the content itself, in particular, to when it is factually unreliable or socially harmful – this is, for instance, content that denies established scientific knowledge, incites hate or promotes fake news. In this way, works that approach content unreliability are focused on detecting harmful content or discussing social implications of its propagation. In Chapter 5, we discuss that with our experimental analysis on YouTube.

2.1 Definitions

In Chapter 3, we analyze reliability over the crowdsourced reviews. In this Chapter, our definition of unreliability relies on time, i.e., how stable ratings are on time. In this way, observing temporal dynamics, we breakdown components to quantify and understand different actors on the change (unreliability). In the literature, although authors are not explicitly referring to the unreliability of crowdsourced reviews as we are, several works present the temporal dynamics on online ratings (McAuley and Leskovec, 2013; Zhang et al., 2014) and some other works also attempt to reason their variability (Duan et al., 2008; Engler et al., 2015; Yang et al., 2012) .

In Chapter 4, we analyze the impact of Recommender System regarding polarization, the spectrum of information made available to users. We consider three dimensions on which polarization effects can be observed, and we analyze them analyzing changes in time of the user's preference vector. The first one, intensity, capture the idea of amplification, i.e., if the recommender system polarizes the user by reinforcing its own opinion and making it stronger. We measure intensity by computing the norm over the user's preference vector. The second polarization dimension, simplification, regards the diversity of the user opinion as a whole, i.e. if the recommender system polarizes the user by strengthening just some narrow set of recommendations. We measure simplification using the entropy over the user's preference vector. The last dimension, divergence, concerns homogenization, i.e., if the recommender system polarizes users by turning them more similar to each other. We measure divergence by averaging the correlation coefficient among pairs of user's preference vectors. The authors in (Dandekar et al., 2013) consider that polarization is a property of an opinion formation process instead of a property of a state of the network. They establish that an opinion formation process is polarizing if it results in an increased divergence of opinions. They measure it in terms of the network disagreement index. Our definition of polarization is also dynamic and applied to the effect of the recommender system on transforming the user's preference over time. We, however, analyze polarization over more dimensions.

In Chapter 5, we analyze reliability over a recommendation system. In this Chapter, our definition of reliability regards the source of information that we categorized as extreme, neutral, and reliable. The definition of reliable information used by authors in (Andrew Guess, 2018) regards "fake-news" – new article (primarily political) that couldn't be credible by fact-checkers. Other articles such as (Bergen, 2019; Lewis and McCormick, 2018; Nicas, 2018; Tufekci, 2018), refer to unreliable information relating to content promoting either misinformation, political extremism, or hate content. This former reference is more aligned to our definition.

2.2 Crowdsourced Reviews

2.2.1 Reliability: Temporal Dynamics

One of the biggest challenges when evaluating the reliability of crowdsourced reviews (this is, assessing reviews real value) is due temporal dynamics. Many previous studies have looked at temporal dynamics in online reviews, but to the best of our knowledge, we are the first that addresses the reliability of online ratings considering the role played by the complete set of factors listed in Section 3.2.

One starting point for our analysis in Chapter 3 is (Koren, 2010), which proposes a recommender system based on collaborative filtering that incorporates temporal dynamics, and splits prediction score between various factors. The authors in (Liu et al., 2017) present a temporal rating model that additionally incorporates review text; we focus just on review scores as a function of time.

McAuley and Leskovec propose a latent factor recommender system in (McAuley and Leskovec, 2013) that models user development caused by the consumption of products over time. They show the role of user experience and expertise through analysis of beer, wine, food, and movie reviews; however, we do not find a significant impact of user evolution in our study. The authors in (Zhang et al.,), (Liu et al., 2010), and (Xiong et al.,) also model temporal dynamics as a strategy to improve recommendation accuracy, and use models similar in spirit to our model; however their purpose is not to understand ratings dynamics. Likewise, the authors in (Li et al., 2017) study how positive and negative movie reviews change over time and propose a recommender system model that takes into account time-varying and temporal effect of positive and negative reviews for future behavior.

While all of these studies propose new methods for improving recommendations, none seeks to understand a broad set of factors underlying the evolution of rating dynamics observed in practice such as platform differences or population shift as we did in Chapter 3 .

The authors in (Godes and Silva, 2012) analyzed the evolution of online ratings over sequence and time for a book ratings dataset. They show that, on average, ratings in sequence and time decrease, although there are distinct dynamics processes occurring. Although they provide some explanations for those dynamics processes their analysis is limited to a specific platform and item type.

Finally, we point out some studies that look at the dynamics of online reviews with focus on some particular correlations. Tha authors in (Duan et al., 2008) model the positive feedback mechanism between online word-of-mouth (WOM) and retail using a movie dataset. The authors in (Engler et al., 2015) create a model to understand online product ratings from a consumer perspective. The compare evaluations of products from consumer magazine and online ratings and observe that besides product quality, online ratings reflect customers satisfaction. The authors in (Salganik and Watts, 2009) analyzed the role of social dynamics in cultural markets. In a similar perspective (Li and Hitt, 2008) and (Yang et al., 2012) analyze in online systems the effect of conformity or social influence bias – the inclination to conform to the observed norm of a community. Furthermore, (Krishnan et al., 2014) proposes a recommender system that mitigates this conformity effect while (Liu et al., 2016) a system to embrace it. Although they model some the temporal dynamics of some effect on online ratings, their analysis does not provide a general understanding of factors affecting those dynamics.

2.3 Recommender Systems

2.3.1 Polarization: Opinion Formation Dynamics

Although a number of papers have addressed the individual components of the dynamics between recommender systems and user opinion formation, or the social implications of these dynamics, few of them have addressed the problem in whole.

Initially, we review three works that have addressed polarization in recommender systems by studying the dynamical system composed by both.

The authors in (Dandekar et al., 2013) explore the causes of polarization by studying an opinion formation process based on averaging of user opinions. They show that the opinion in the group polarizes when users responses have biased assimilation; this is, the response reinforces their own current opinion. Additionally, they expand the opinion formation process analysis to a recommender system model analyzing the biased assimilation response. This work is discussed in detail in Chapter 4, and it was used as a starting point for our link-based analysis. As we cover in Section 4.3 our results generalize that work and show that in many realistic settings, the authors' conclusions therein do not apply.

Bakshy, Messingh and Adamic carry out an empirical analysis of the ideological diversity of news exposure on Facebook in (Bakshy et al., 2015). They measure the diversity of news (through political alignment) shared between friends and measure the potential for cross-cutting – i.e. how much a user from a given political alignment (conservative, moderate or liberal) is exposed to news of a different political alignment. They concluded that the individual's social network itself is the most important factor in limiting their exposure to diversity. However, is the user's choices about what to consume (i.e. which links to click) more than the newsfeed ranking algorithms that contribute to the news diversity consumed. Although (Bakshy et al., 2015) recognized the dynamics between news feed (recommender system) and users, their analysis focuses only on the user's exposure to diversity when user opinion (political affiliation) is held constant in their model.

Koren in (Koren, 2010) proposes a recommender system based on collaborative filtering that incorporates temporal dynamics. Using his recommender system, he splits the prediction score between factors dependent or independent of the interaction of users and items and performs an analysis of rating drifting on a Netflix dataset. Although (Koren, 2010) studies the dynamics of recommender systems and how much the interaction of user and items can affect the prediction score over time, his analysis is focused on the recommender system itself and doesn't explore the user's opinion formation.

Next, we turn to work that addresses individual aspects of polarization in recommender systems.

A number of papers have addressed situations in which there is a narrowing of access to information by users in online systems. Although these works cover polarization, their focus has primarily been on personalization and recommender engines instead of its dynamics as explored in this thesis. This effect has been dubbed a "filter bubble" by Pariser in (Pariser, 2011). Following this line, researchers have analyzed interactions between users and recommender systems to detect filter bubbles. The authors in (Hannak et al., 2013) and (Kliman-Silver et al., 2015) explore the factors that trigger personalization on results of search queries. Furthermore online price discrimination was exposed by (Hannak et al., 2014) and online ad delivery discrimination was exposed by Sweeney in (Sweeney, 2013). These are examples in which information content is filtered and throttled in a fashion that is undesirable.

The authors in (Zuiderveen Borgesius et al., 2016) investigate the different types of personalization in communication and they claim that at present, there is no empirical evidence that warrants any strong worries about filter bubbles, however, they agree that this could be a future problem if personalization technology improves.

Another line of work related to the filter bubble phenomenon tries to understand the role played by recommender algorithms in decreasing recommendations diversity and propose solutions to avoid that this limit the user experience.

Knijenburg et al in (Knijnenburg et al., 2016) claim that to solve the filter bubble problem it is necessary to build a recommender system with a different goal than simply recommending good items. The authors of (Knijnenburg et al., 2016) then propose an idea of a new recommender system – named Recommender System for Self-Actualization – that aims to support the users in developing, exploring and understand their unique taste.

Graells-Garrrido et al in (Graells-Garrido et al., 2013) propose a content recommender on Twitter that uses graphical tools and gap indicators to stimulate diversity and connect people of opposite views. Other researchers such as (Maccatrozzo, 2012), (Abbassi et al., 2009), (Zhang et al., 2012) and (Oku and Hattori, 2011) suggest the inclusion of "serendipitous" recommendations in order to promote diversification. While these suggestions are consistent with our observation that "best" recommendations are not always the most beneficial for user opinion formation, they also do not consider how opinions and recommender systems evolve together over time.

Finally, we examine work that takes one step back in respect to our analysis of dynamics of polarization and that covers the process of opinion formation. Prior work in opinion formation has taken several approaches to understand how opinions can evolve in a network according to a given model of information propagation. One line of work examines the direct influence of neighbors and self-beliefs in opinion formation; this is studied in (Bindel et al., 2012) where a repeated averaging model is used to analyze consensus. An alternative approach, taken by (Bhawalkar et al., 2013) uses game theory as opinion formation process. Both works present boundaries for the cost to reach consensus in their models. Finally, in (Mäs M, 2013) Mas and Flache propose a peer-to-peer interaction model that can explain the polarization of opinion with homophily and without negative influence (disliking of dissimilar others). Although these models are concerned with some of the same phenomena as our study (eg, polarization) they do not include a recommender system as an external agent.

2.3.2 Reliability: Content Evaluation

A number studies have explored the impact of recommender systems related to the reliability of content. In this section, we contrast our work with those studies.

Our work in Chapter 5 takes inspiration from recognition in the popular press that YouTube's recommendations can lead to extreme content. Articles such as

(Tufekci, 2018) and (Lewis, 2018) describe the radicalization of YouTube recommendations and discuss the social implications of this effect. In a similar manner, (Chaslot, 2018) argues that using AI for optimizing engagement could discredit the media, and it provides examples of some anti-media content that has been recommended. The article (Bergen, 2019) continues the discussion and presents some of the actions taken by YouTube to address the criticism about unreliability recommendations. The articles (Roose, 2019) and (Kevin Roose, 2019) reinforce that criticism and describe some positioning and action from YouTube to address them. Furthermore, similar criticism has also been applied to Amazon's recommender system in (Diresta, 2019), and (Oram, 2019) discusses the possibility of more democratic balance on news delivery by the new "Apple's News+" service (which is not advertising-driven at present). These articles provide an essential context for our study by highlighting issues, but none performs a quantitative analysis of YouTube's recommendation system. In contrast, we quantify the strength and dynamics of YouTube's "lead away" effect, showing for example that most of its effect takes place within a sequence of just a few recommendations.

In (Nicas, 2018) the author considers the social implication of YouTube recommendations and investigates the outcome of the most frequently returned videos starting from trend searches. The author also performs an empirical exploration of YouTube recommendations. Similar to our results, that study finds that YouTube's recommendations often lead users to channels that feature conspiracy theories, partisan viewpoints, and misleading videos. However, that work looks at a much smaller dataset with a simpler overall experimental design; by studying a much larger dataset we provide greater robustness of results. Most importantly, it does not explore the trade-off between recommendation properties and privacy, nor does it analyze time dynamics, the impact of YouTube policy changes, nor the relationship to query topic.

The authors in (O'Callaghan et al., 2015) investigate the recommendation of extreme-right videos on YouTube by using a content categorization schema. Like our study, that work notes how quickly the YouTube recommender system can deviate from reliable content. However, that work focuses on one specific niche of the content spectrum (extreme-right) and the discovery of its ideological bubbles. Our work adopts a more extensive notion of extreme content and consequently provides a broader understanding of the "lead away" effect of YouTube recommendations. Further, other work also covers the discovery of ideological bubbles with harmful social impact on YouTube, but without examining YouTube recommendations. In this regard, the authors in (Sureka et al., 2010) propose a framework to discover hate videos on YouTube and, authors in (Bermingham et al., 2009) perform sentiment analysis on comments to identify online radicalization. These studies can complement ours as tools for channel classification.

Moving beyond YouTube, the authors in (Le et al., 2019) analyze personalization on Google News based on a user's browsing history. Their study also analyzes how revealed user information can affect the outcome of recommendations, but they are focused only on the political echo chamber dimension, and they do not consider varying degrees of privacy as we do.

Chapter 3

Unraveling the Dynamics of Online Ratings

3.1 Introduction

One of the ways that the Web has revolutionized society is through crowdsourced reviews. Almost any situation in which alternative choices may be evaluated is now supported by one or more review systems that record experiences and ratings that users have provided for items of interest.

From the standpoint of the review user, the value of a review system is to allow the user to assess the perceived quality of various alternatives before making a decision. However, there is considerable evidence that online reviews show considerable temporal dynamics (McAuley and Leskovec, 2013; Zhang et al., 2014), so an important question concerns how to understand the dynamics of item reviews before using them to make decisions.

In this chapter we seek to understand how item ratings change over time and what factors affect those changes. This is a complex question because there are many dimensions that can play a role in review dynamics. Of course, ratings may shift because the popular perception of an item is actually changing within the user population. However, many more factors come into play. For example, ratings can be affected by shifts in the nature of the population of users providing ratings. They can be affected by closed-loop effects in which previous ratings influence the set of users that are interested in and subsequently review the item. Ratings shifts can occur for some items in a manner that is different from other items. Furthermore, the dynamics of ratings can differ among different ratings platforms – even for the same set of items.

In this chapter we show how to tease apart all of these effects, characterize them, and quantify their relative importance. Our goal is to form an integrated view of how the interplay of these effects ultimately determine the changes in item ratings over time. To do so, we make use of a variety of datasets, chosen for their ability to explore all of the questions above. To study platform effects, we study the dynamics of movie ratings across three major ratings platforms; and to study item category effects, we study various item categories on a single platform. We use clustering to distinguish items showing different rating dynamics on the same platform. And within a given platform and item category, we fit a nonlinear model that allows us to distinguish factors such as changes in user population, changes in user behavior, intrinsic changes in perceived item quality, and closed-loop interactions between previous ratings and changes in user population. This latter factor essentially captures the impact of the ratings platform as a recommender system.

Our multi-platform, multi-factor study goes beyond prior work by considering a much broader set of factors than previous studies. Using our methods, we show that there are consistent differences in rating dynamics that depend on the nature of the rating platform. These differences are not due to different sets of items being rated on different platforms – they persist even when looking at the same set of items on different platforms. We also show that on each platform, there are understandable shifts in the kinds of users that rate an item over time, and that in each case this population shift makes sense due to the nature of the platform. We show that there are consistent general trends in how perceived item quality changes over time, which are understandable in light of past studies. And we show that recommender systems play a role in affecting rating dynamics on some platforms, but not others, in a way that correlates with the nature of the rating platform.

3.2 Factors Affecting Rating Dynamics

Our goal is to form a holistic picture of the forces that combine to determine online rating dynamics. In particular, we seek to understand how the following factors interact in shaping online ratings:

Platform Characteristics. We consider first, does the platform explicitly support item sales, or is it purely informational? And second, does the platform provide a recommendation system as a service, or does it merely display ratings?

User Population. We want to evaluate whether their are different user types, and if so whether the balance among those types changes over time, and how those shifts affect ratings dynamics.

Item Perception. We seek to quantify the extent to which the popular perception of an item is shifting over time. This can reflect a shift in tastes in the population at large, or a tendency for a less-appreciated item to become better appreciated by the population over time.

Item Type. We seek to understand whether different types of items show different dynamics, and why. We also seek to understand the prevalence of non-trivial dynamics, i.e., the proportion of items within a category that typically show detectable dynamics over time, as opposed to the proportion of items whose ratings are approximately unchanging.

Closed-Loop Effects. Finally, we are interested in the extent to which online ratings or recommendations affect the set of users that subsequently consume an item, leading to shifts in dynamics of future ratings. This tells us the impact of "tuning" between items and the users that consume and rate the items, a tuning that is induced by recommendations.

To separate and evaluate these effects, we use the data and methods described in the next section.

3.3 Methods

In order to effectively disentangle all of these effects, we use a combination of carefully chosen datasets, unsupervised learning in the form of clustering, and supervised learning in the form of a model fitted to our various datasets.

3.3.1 Data

We make use of the following datasets to help distinguish the five factors above:

Movie Tweetings. This dataset is collected from well-structured movie evaluation tweets on Twitter from 2013 to 2017 (Dooms et al., 2013). This dataset represents a platform in which there is no explicit recommendation system, and there is no commercial entity providing the reviews for the purpose of commerce.

We selected a relatively dense subset of this dataset, namely movies that have at least 10 ratings and users that have at least 5 ratings. We denote this dataset `MT`. This dataset has 15632 users, 5780 movies and 521214 ratings. We centered the ratings in `MT` by rescaling them from [1:10] to [1:5].

Rotten Tomatoes. This dataset was crawled from the Rotten Tomatoes website (Fandango, 2016) in late 2016, which we denote as `RT`. This dataset represents a platform in which there is a known distinction between two user types: *critics,* and *general users.* Like `MT` there is no explicit recommendation system or commercial role for the platform.

From the entire dataset we also selected the subset consisting of movies that have at least 10 ratings and users that have at least 5 ratings. The resulting dataset has 165585 users, 12122 movies, and 4845884 ratings. We centered the ratings in RT by rescaling them to the range [1:5].

Amazon. This dataset contains product reviews from Amazon spanning from May 1996 to July 2014 (McAuley et al., 2015) and (He and McAuley, 2016). This dataset represents a platform in which there is an explicit recommendation system that makes personalized purchase suggestions to users. The platform also has a commercial role in support of sales in the Amazon store. Furthermore, the Amazon dataset contains items from multiple categories. In addition to movies, we use it to study electronics, home goods, CDs, mobile apps, and ebook (Kindle) titles. From the Movies and TV category, we first disambiguated movies names, including merging movies available in different media such as DVDs and BluRay which appeared as separate products. Next we select a dense subset of movies that had at least 5 reviews. We denote this dataset AZ, and it has 1957899 users, 53633 movies, and 4291173 ratings.

From the other categories, we selected their 5-core dataset - the dense subset of items with at least 5 reviews and users with at least 5 reviews. The resulting datasets are: Electronics (AZ-Ele) and with 192401 users, 63001 items, and 1689129 ratings; Home and Kitchen (AZ-Hom) and with 66518 users, 28237 items, and 551656 ratings; Kindle Store (AZ-Kin) and with 68222 users, 61933 items, and 982197 ratings; Apps for Android (AZ-App) with 87267 users, 13209 items, and 752832 ratings; and CDs and Vinyl (AZ-CDs) with 75256 users, 64443 items, and 1097555 ratings. Note that in what follows, AZ refers to Amazon movies, while the other Amazon categories have specialized names.

3.3.2 Modelling Temporal Dynamics

Definitions

In each application of our model, we consider a dataset having n users and m items. Items are objects over which the user provides a rating, e.g., movies. Each rating has an associated timestamp t (in units of days), and we denote a rating provided by user u for item i at time t as $r_{ui}(t)$. All ratings range from 1 (worst) to 5 (best).

For each rating, we define an associated *system time* which is the time since the item first appeared in the system. That is, if $t_0^{(i)}$ is the timestamp of item i's first recorded rating and t the timestamp of a given rating $r_{ui}(t)$, the system time for that rating is $t_s = t - t_0^{(i)}$.

In presenting our results, we are primarily concerned with *item progression.* This is defined as the index of where a review falls in the ordered set of reviews for an item. So item progression from 0 to 99 reflects the first 100 reviews of an item in order (regardless of how much real time elapsed between the first and last reviews in the sequence).

Model

To separate the factors at work in a single dataset, we fit the data to a predictive model we call `timeSVD--`. This model is a simplified version of the `timeSVD++` for collaborative filtering as proposed in (Koren, 2010) .

To model a rating $r_{ui}(t)$, `timeSVD--` incorporates three kinds of information. First, it uses properties of the user u: a term capturing the user's time-invariant average rating (bias), and a term capturing the evolution of the user's average rating over time. Second, it uses properties of the item i: a term capturing the item's time-invariant average rating, and a term capturing the evolution of the item's average rating over time. Finally, it incorporates latent factors for both the user and item, whose inner product models the personalization of the item to the user. This latter factor is essentially a matrix-factorization approach to personalization (as reflected by the 'SVD' in the name of the model).

Specifically, `timeSVD--` is parameterized as follows:

μ	Global mean of all ratings
b_i	Time-invariant bias (average rating) of item i
$b_{i,Bin(t)}$	Time-varying bias of item i at timebin $Bin(t)$
b_u	Time-invariant bias of user u
$\alpha_u \mathrm{dev}_u(t)$	Time-varying bias of *user* u
q_i	k-dimensional latent factor of item i
$p_u(t)$	Time-varying k-dimensional latent factor of user u

The model reflects the assumption that user preferences may change over time ($p_u(t)$) while item features are time-invariant (q_i).

The `timeSVD--` model is then:

$$r_{ui}(t) = \mu + b_i + b_{i,Bin(t)} + b_u + \alpha_u \mathrm{dev}_u(t) + q_i^T p_u(t) \tag{3.1}$$

`timeSVD--` incorporates various strategies to capture time evolution of model components without unduly expanding the set of parameters to be learned. In the case of item bias, time is discretized into bins of seven days. For time-varying user parameters, the model fits a symmetrized polynomial:

$$\text{dev}_u(t) = sign(t - t_u)|t - t_u|^\beta$$

where t_u is the mean date of rating of *user u*. This function is used in time-varying user bias as well as in the time-varying user latent factor:

$$p_{u\ell}(t) = p_{u\ell} + \sigma_{u\ell}\text{dev}_u(t) \quad \ell = 1, \ldots, k$$

In the rest of the chapter, we will refer to $q_i^T p_u(t)$ as the *interaction* score between u and i, and the rest of the terms in (3.1) as the *baseline* score between u and i.

Learning the Model

We train `timeSVD--` on each dataset using system time t_s as the value of t for each rating. To learn model parameters we apply stochastic gradient descent to a risk function incorporating a regularization to (3.1):

$$f(\theta) = \textstyle\sum_{\text{all ratings}}(r_{ui}(t) - (\mu + b_i + b_{i,Bin(t)} + b_u + \alpha_u \text{dev}_u(t) + q_i^T p_u(t)))^2$$
$$+\gamma(\textstyle\sum_i(b_i^2 + ||q_i||^2 + \sum_{Bin(t)} b_{i,Bin(t)}^2) + \sum_u(b_u^2 + \alpha_u^2 + ||p_u||^2 + ||\sigma_u||^2))$$

We set model hyperparameters γ and β by cross-validation.

Clustering

Within a particular dataset, we expect different items to show different dynamics over time. In order to efficiently separate items by the properties of their ratings dynamics, we use a clustering algorithm well-suited to work on timeseries: `k-Shape` (Paparrizos and Gravano, 2016). To study factors at work for different kinds of items, we apply both `timeSVD--` and `k-Shape,` and take averages of the `timeSVD--` results over clusters identified by `k-Shape`.

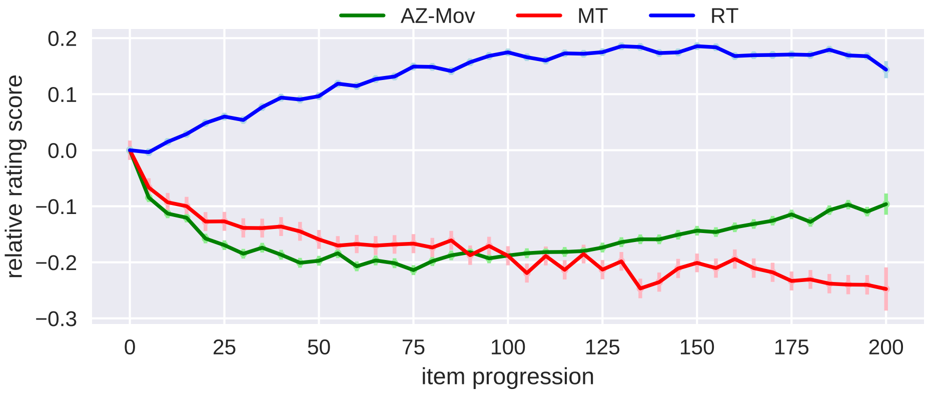


Figure 3·1: Relative ratings progression

3.4 Analysis

We divide our analysis into two parts: first we characterize the range of observed phenomena in review dynamics, and then we decompose those phenomena to gain understanding of how they arise.

3.4.1 Characterizing Ratings Dynamics

Our basic tool for studying review dynamics is the *relative rating score.* This is the average value of ratings on a daily basis, offset by a constant that makes the first set of average ratings equal to zero. We call the average value of the first item ratings the *initial score* and the average across the study period (usually 200 reviews) the *average score.* We focus on *item progression* which, as described above, is the ordered sequence of reviews for an item.

Throughout our analysis consider only movies with at least 200 reviews, and analyze the first 200 reviews. This means that the set of movies contributing to each average rating is not changing over time.

How do item ratings change over time? We start by addressing this basic question in Figure 3·1. This figure shows that each dataset shows distinctive behavior.

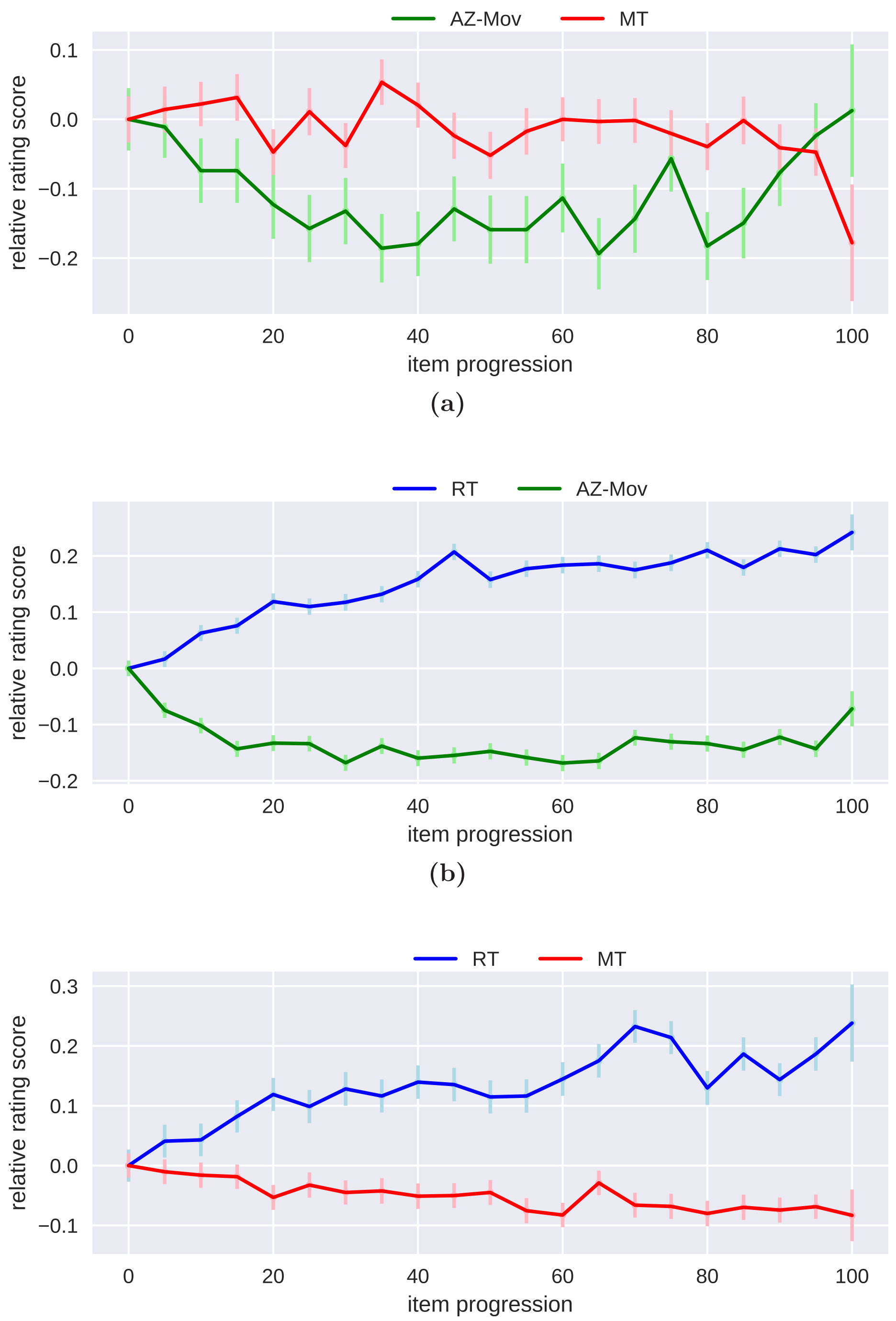
The RT dataset shows a generally increasing trend; the MT dataset shows a generally decreasing trend; and the AZ dataset shows a trend that first decreases, and then increases.

Are platform-specific ratings dynamics consistent? One possible explanation for the platform-specific differences in rating dynamics shown in Figure 3·1 could be that they are due to the fact that the set of movies rated on each platform is different. First, we show that differences shown in Figure 3·1 in platform-specific dynamics are *not* due to the different sets of movies rated.

For each pair of platforms, we select the set of movies that are rated at least 100 times on both platforms (we use a smaller window of 100 reviews to increase the size of the sets being analyzed). We match movies based on title and year (where available), discarding any cases in which duplicate matches occur. Figure 3·2 shows the item progression for movies in common between each pair of datasets, and that in each case, platform-specific trends are preserved. Specifically, Figure 3·2a shows the 127 movies in common among AZ and MT, Figure 3·2b shows the 1451 movies in common among RT and AZ, and Figure 3·2c shows the 387 movies in common among RT and MT.

In each case, the platform-specific trends shown in Figure 3·1 a preserved (although due to the smaller dataset sizes, there is more variability and trends are correspondingly weaker in some cases.) We conclude that the differences shown in Figure 3·1 are consistently present when studying the same sets of movies on different platforms.

We also note that the relationships between initial and average scores across platforms are preserved when restricting attention to common movies, with AZ $>$ MT in Figure 3·2a ((4.24, 4.12) $>$ (3.87, 3.86)), AZ $>$ RT in Figure 3·2b ((4.29, 4.16) $>$ (2.93, 3.08)), and MT $>$ RT in Figure 3·2c ((3.59, 3.54) $>$ (2.87, 3.00)).



(a)

(b)

(c)

Figure 3·2: Common movies: relative ratings progression. Subfigures are: (a) AZ-Mov and MT: 127 common movies; (b) RT and AZ-Mov: 1451 common movies; (c) RT and Movie MT: 387 common movies.

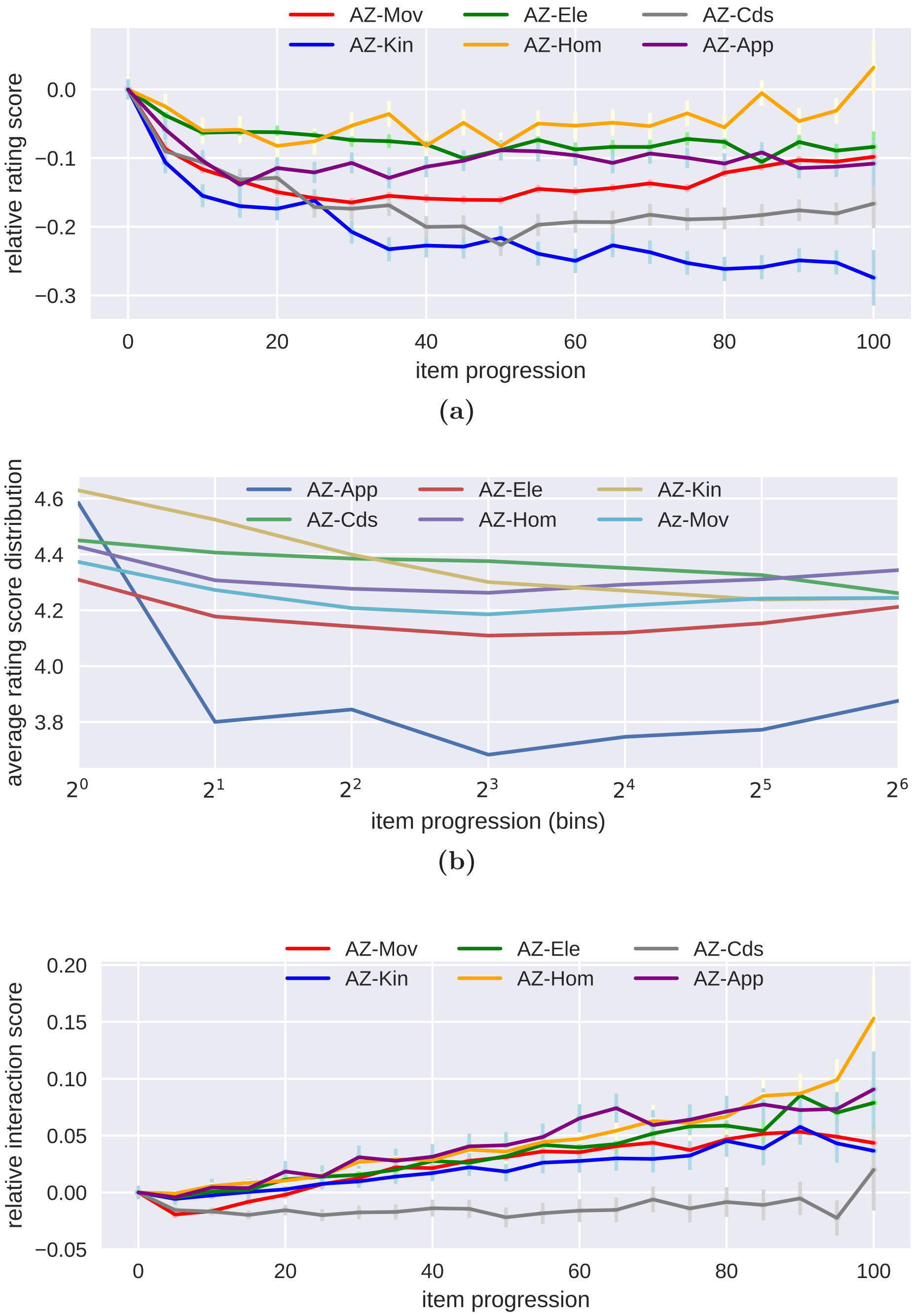
Another way to assess whether platform-specific ratings dynamics are consistent is to ask whether the same dynamics are seen across multiple item categories on a given platform. To confirm this, we look at relative ratings scores across the six categories of Amazon data, shown in Figure 3·3a . This Figure shows that the general behavior of declining followed by increasing ratings is widespread across most of the item categories on the Amazon platform.

The above results suggest that the platform-specific ratings dynamics we observe are not due solely to differences in items rated on the different platforms, but rather that these effects are relatively consistent.

Do all items change in the same way within each platform? A final characterization question concerns how the platform-wide effects seen in Figure 3·1 are produced from the individual contributions of each movie. We explore this question by clustering the movies individual item progressions using the `k-Shape` algorithm (Paparrizos and Gravano, 2016), and studying cluster-wide averages. We use a default of five clusters in each case, which we observe to balance clear separation of classes against noise introduced due to small samples.

Figure 3·4 shows the results of this clustering for all platforms. The Figure shows that the characteristic dynamics on each platform are not always present for all movies. In the case of `AZ`, the characteristic decrease/increase pattern is primarily present in a cluster 1, comprising about 12% of all movies. The other clusters primarily show a simpler decreasing trend. In the case of `RT`, the characteristic increase is primarily present in clusters 0 and 4, comprising about 45% of all movies. Finally, in the case of `MT`, in general all movie clusters show the platform's characteristic downward trend, with the strongest trends in clusters 0, 2, and 3.

We conclude from Figure 3·4 that not all items are showing strong dynamics in each dataset and that, furthermore, dynamics are not occurring uniformly in each. As

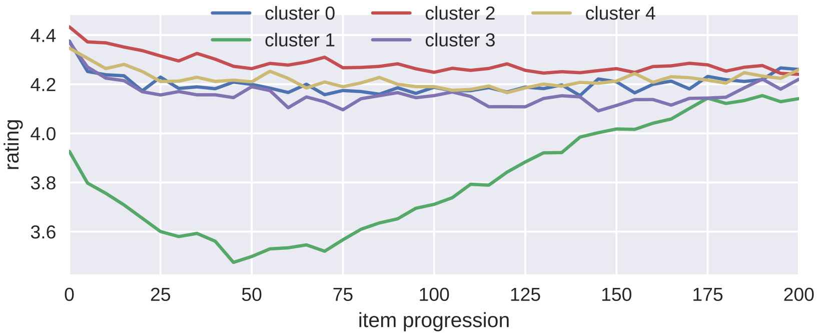


(a)

(b)

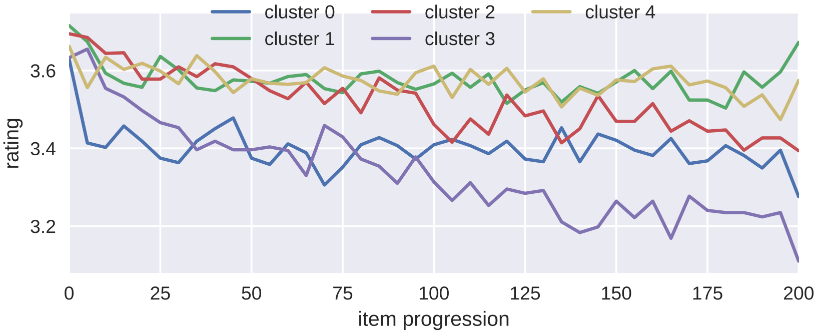
(c)

Figure 3·3: Relative progression across Amazon categories. Subfigures are: (a) Relative Ratings Progression; (b) Amazon early user effect; (c) Amazon Interaction Score - all items.



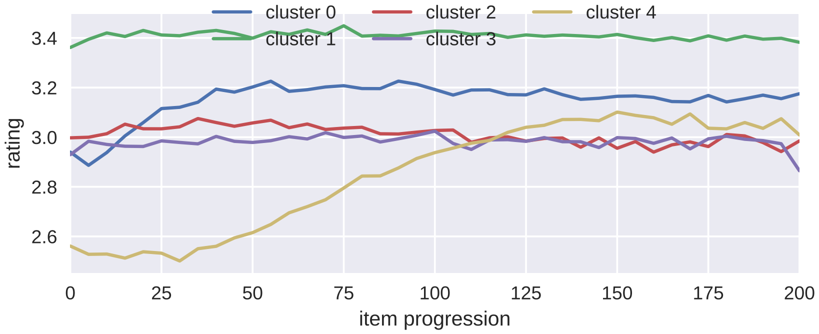
cluster	size
0	834
1	502
2	1127
3	739
4	948
All	4150

(a)



cluster	size
0	87
1	116
2	127
3	109
4	141
All	580

(b)



cluster	size
0	1472
1	1535
2	1073
3	1020
4	1450
All	6550

(c)

Figure 3·4: Relative ratings progression by cluster. Subfigures are: (a) Amazon (`AZ`); (b) Movie Tweetings (`MT`); (c) Rotten Tomatoes (`RT`).

a result, in what follows we will generally distinguish between “large effect” movies (AZ cluster 1, MT clusters 0, 2, 3, RT clusters 0,4) and “small effect” movies (movies in the remaining clusters).

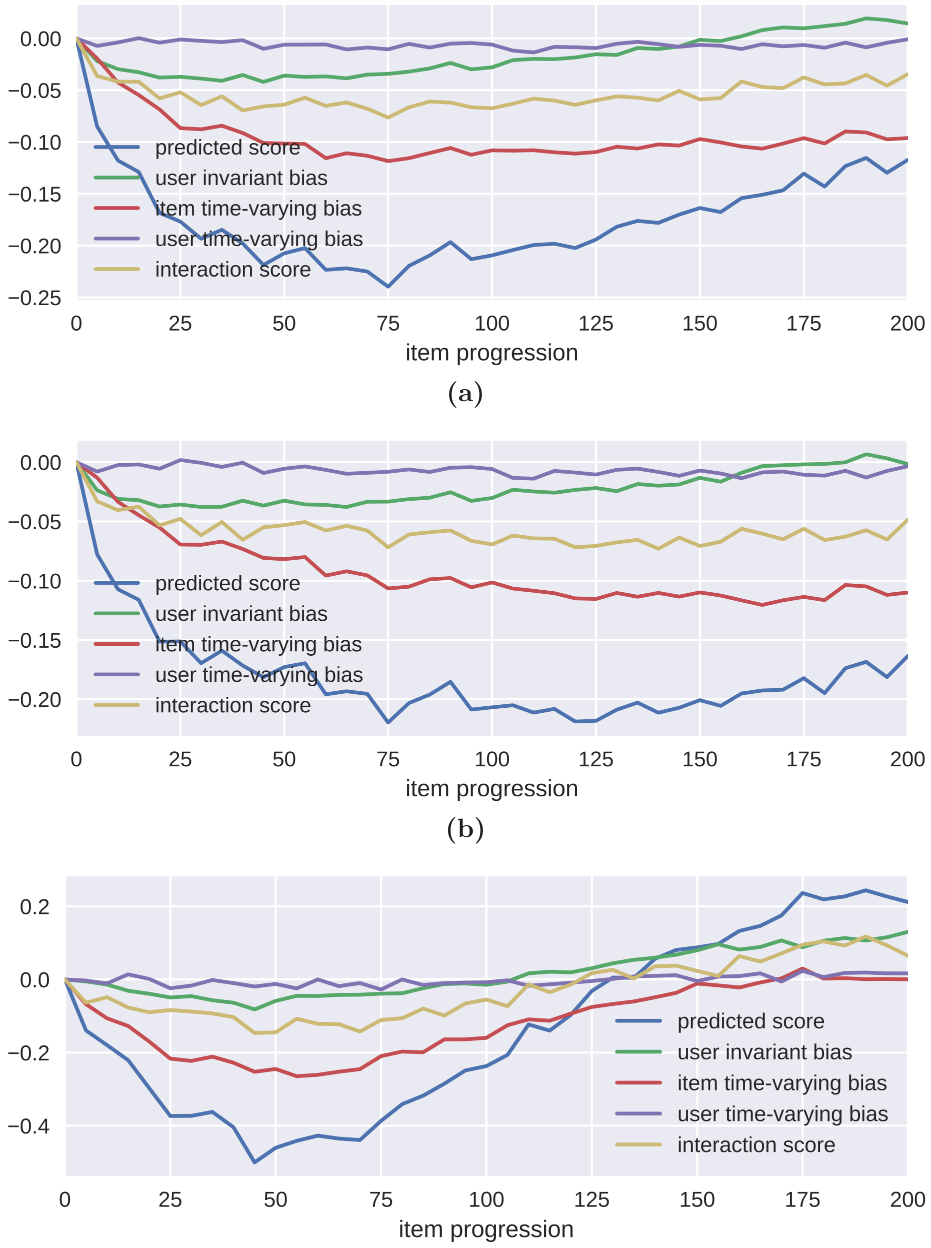
3.4.2 Decomposing Ratings Dynamics

To develop an understanding of the forces driving the effects seen in Section 3.4.1, we decompose ratings using timeSVD--. The components of the model bear direct relationship to various factors of interest as described in Section 3.2. In particular, we can study the impact of the user population by looking at the user time-varying and invariant components of the model ($\alpha_u \text{dev}_u(t)$ and b_u), we can study the impact of item perception by studying the item time-varying and invariant components of the model ($b_{i,Bin(t)}$ and b_i), and we can study the impact of closed-loop effects by studying the model's interaction score ($q_i^T p_u(t)$).

What are the main model factors affecting ratings dynamics? We start by decomposing the three datasets according to our model, and according to movie type (small effect vs. large effect as described above). The results are shown in Figure 3·5, Figure 3·6 and, Figure 3·7.

We start first with Figure 3·5a, Figure 3·6a, and Figure 3·7a which show relative contributions of factors, respectively, for AZ, MT, and RT. There are a number of high-level observations. First, user invariant components and item time-varying components are the largest and primary contributors to ratings dynamics. Furthermore, the only platform in which interaction score shows significant dynamics is AZ.

Figures 3·5c, 3·6c, and 3·7c show the corresponding breakdowns for the large-effect movies, and the results there confirm the conclusion that user invariant and item time-varying components are the main contributors to the respective platform dynamics. (Figures 3·5b, 3·6b and 3·7b show the small-effect movies – note the difference in scale on the y-axes).

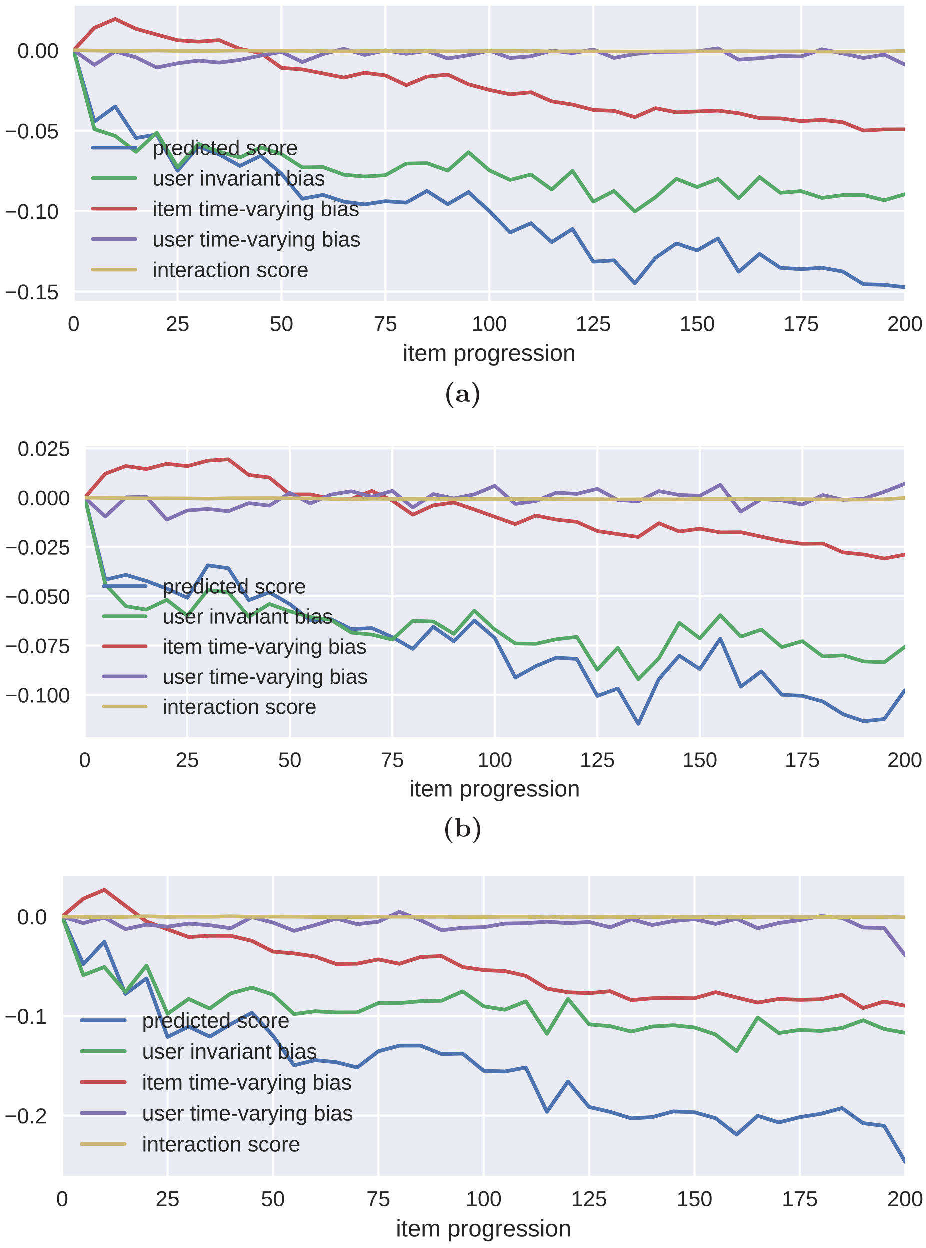


(a)

(b)

(c)

Figure 3·5: Amazon (`AZ`) Relative `timeSVD--` Components Progression. Subfigures are: (a) `AZ` all movies; (b) `AZ` small effect movies; (c) `AZ` large effect movies.

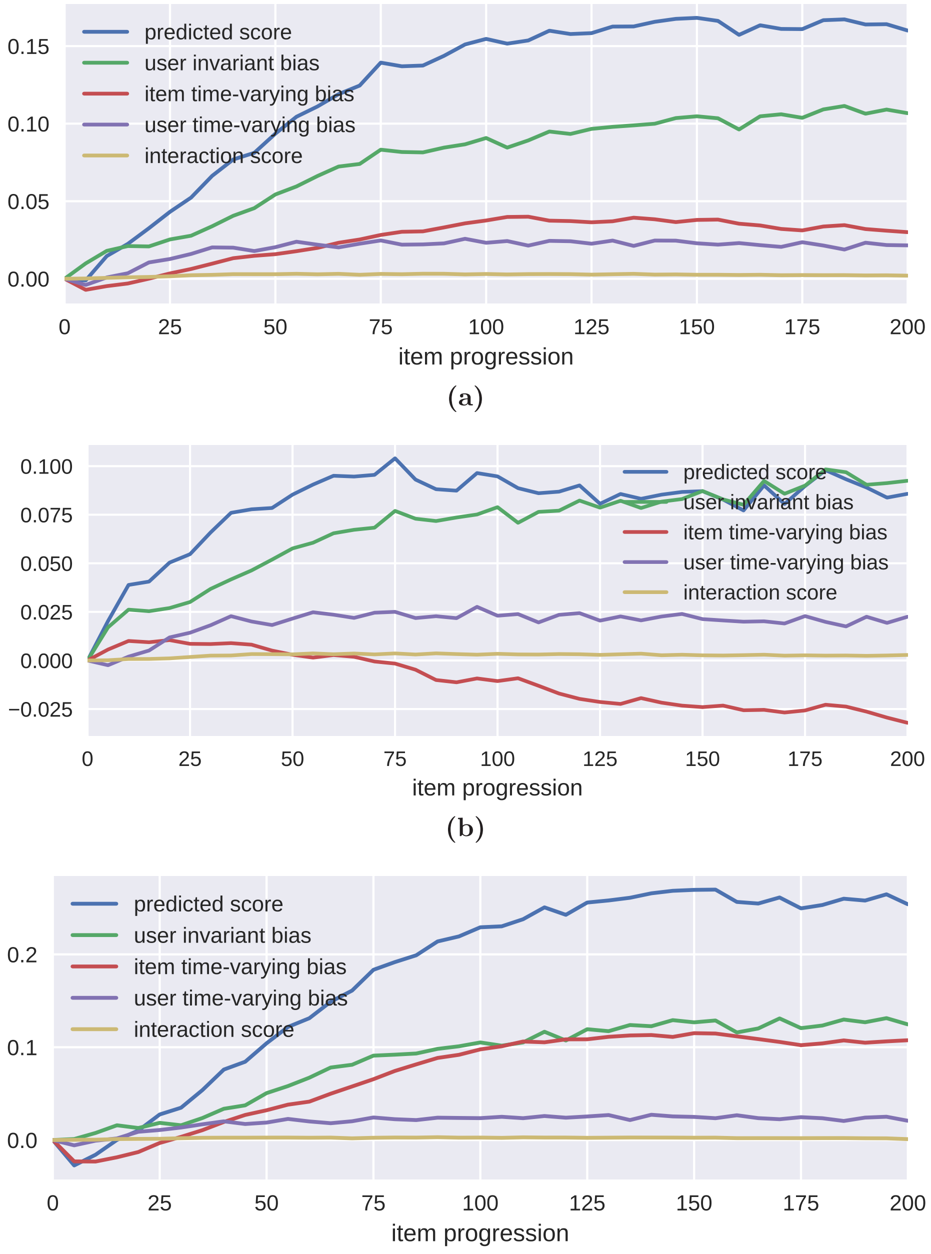


(a)

(b)

(c)

Figure 3·6: Movie Tweetings (`MT`) Relative `timeSVD--` Components Progression. Subfigures are:(a) `MT` all movies; (b) `MT` small effect movies; (c) `MT` large effect movies.



(a)

(b)

(c)

Figure 3·7: Rotten Tomatoes (RT) Relative `timeSVD--` Components Progression. Subfigures are: (a) RT all movies; (b) RT small effect movies; (c) RT large effect movies.

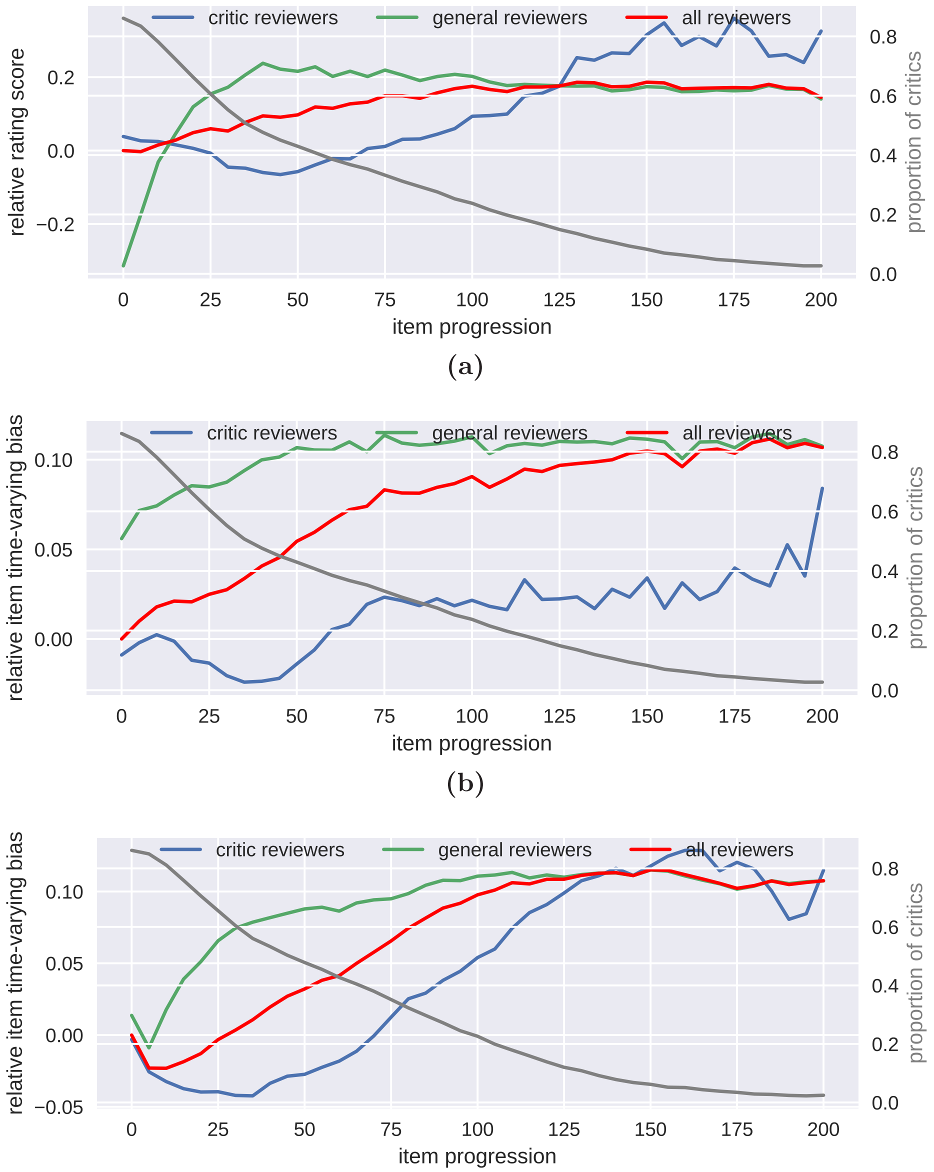
We now explore each of the factors in turn.

How do users contribute to rating dynamics? We first examine the role of users in rating dynamics. We note from Figure 3·5, Figure 3·6 and, Figure 3·7 that the user time-varying component (purple line) does not show significant contribution to rating dynamics, but the user time-invariant component (green line) does show significant contribution. This means that while users individual ratings averages are not changing over time, users' contribution to changes in ratings are nonetheless significant. In other words, *changes in the user population* – in a consistent way – are a major driver of ratings dynamics (on all three platforms).

For AZ the contribution of changes in user population (green line) reflects the overall platform pattern of initial decline followed by increase. This component contributes about 50% of the overall change at the end of the 200 review period. For MT the contribution of changes in user population has a decreasing trend of similar range for the whole dataset analysis (Figure 3·6a) as well both subsets (Figures 3·6b and3·6c). This also covers above 50% of the relative changes in ratings for the large-effect set (Figure 3·6c). For RT, we also see that changes in user population play a significant role, contributing about 50% of the change in the large effects subset.

To understand how this significant shift in user population comes about, we turn to the RT dataset. In that dataset, we have the advantage that users are classified as either (professional) critics or general reviewers. We use this classification to achieve a better understanding of role of user population in rating dynamics.

Figure 3·8 breaks down relative ratings score according to user type in RT. In each plot of that figure the blue line represents the critic's reviewers contribution, the green line represents the general's reviewers contribution, the red line the general contribution of all reviewers. The grey line (with y-axis scale on the left side) represents the proportion of critics that are reviewers contributing to the average across movies.



(a)

(b)

(c)

Figure 3·8: Rotten Tomatoes: population change between critic reviewers and general reviewers. Subfigures are: (a) Ratings - All; (b) User time-invariant component - All; (c) Item time-varying component - large.

The proportion (grey line) of critics is the same in all plots.

Figures 3·8a and 3·8b shed considerable light on the user population component of ratings dynamics. It shows that critics are responsible for most of the initial reviews, and that critics on the whole tend to have lower average reviews than general users. (The facts that the user line in each figure drops at the beginning, and the critic lines rise at the end, are due to small-sample effects.) The effect is particularly clear when extracting just the user time-invariant component in Figure 3·8b.

The contrast to `MT` is interesting, because there the user population shift has a decreasing effect on ratings. We note that the `MT` platform is quite different from the other two, because of the absence of a well-defined critic population, as well as the fact that previous ratings of a movie are not as easily accessible. We hypothesize that this means that users whose average ratings are lower will be more likely to review movies later in time.

Overall, this analysis goes a long way to explaining how user population shifts contribute to ratings dynamics. In `RT` it can help explain the entire dynamics of the user time-invariant component of the model. In `AZ` it can help explain the eventual increase in the user time-invariant component; we will explore the initial decrease later in the chapter.

How do items contribute to ratings dynamics? The second significant component exposed by `timeSVD--` in Figure 3·5, Figure 3·6 and, Figure 3·7 is the item time-varying contribution (red line).

For `AZ` in the all movies case (Figure 3·5a) we can observe that the item time-varying component accounts for a substantial proportion of the relative change in ratings score, reaching almost 80% at the end of the progression. For `MT` the item time-varying component always has a decreasing trend – i.e., when the items lose value while aging. We note that this is consistent with previous work (eg, (Godes and Silva, 2012)) showing that, in the absence of other factors, online ratings tend to decline when prior reviews are hard to access or evaluate. The fraction of contribution is considerable – reaching up to 40% of the relative score – in the large effect subset (Figure 3·6c), and it is present across essentially all clusters within the `RT` dataset (Figure 3·4b). Finally, in `RT` the item time-varying component has an increasing trend – i.e. items get a higher score when time progresses – for the all movies case (Figure 3·7a) and for the subset of large effect moives (Figure 3·7c) where it accounts up to 40% of the relative predicted score. The difference in the case of `RT` can also be understood in the context of (Godes and Silva, 2012) due to the presence of a large set of reliable reviews (reviews that are labeled as coming from critics).

Movies with a strong increasing time-varying component are those that show significant improvement in rating over time; they can be thought of as "sleepers" that take time to become well-liked. A more detailed analysis including the separation of the item time-varying component among critics and general reviewers in Figure 3·8c shows that general reviews tend to view "improving" movies earlier in time, while critics tend to view "improving" movies later. This suggests that users are quicker to identify "sleeper" movies and that critics follow. Overall, our analysis shows that on a platform like Twitter, movies with declining ratings over time are more likely to accumulate subsequent reviews than on platforms like Amazon and Rotten Tomatoes, where previous reviews are more accessible and more easily interpreted.

Why do ratings initially decline on the Amazon platform? One of the striking properties of ratings on the Amazon platform is the initial decline followed by subsequent increase. Figure 3·1 shows this effect, Figure 3·4a shows that it primarily derives from about 12% of all movies (although most movies show an initial decline) and Figure 3·5c shows that the effect has contributions from both item time-varying and user time-invariant components. This behavior constrasts starkly with the case

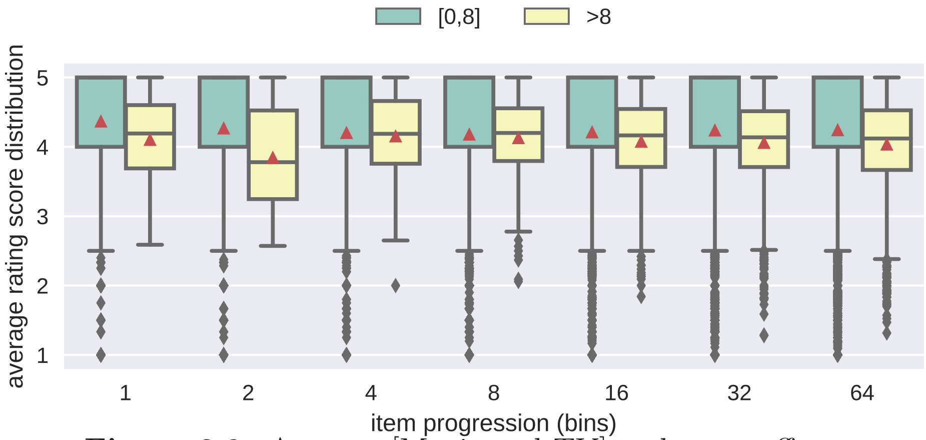


Figure 3·9: Amazon [Movie and TV] early user effect.

for MT and RT.

In investigating this we note that AZ has a numerous quantity of users that just provided a small number of reviews; this, combined with the fact that Amazon is an e-commerce platform, raise the questions of whether initial reviews are intentially inflated in some way. This could be a strategy to attract buyers when a product is first introduced.

We investigated this hypothesis by analyzing the users in the AZ dataset. We conjecture that if large numbers of early reviews were artificially inflated, then there should be a subpopulation of users who are providing almost exclusively early reviews for items.

Hence, fo each user, we compute their average movie rating time (i.e., how early in the item progression time the user provides a review), the user's average rating score, and the number of reviews that that user provided. We summarize the results in Figure 3·9.

In the Figure 3·9, we show the distribution of average rating score of a user versus the average item progression time for that user's ratings. We separate users that contributed less than eight reviews from those that gave more than eight. The figure shows that users that proffered less than eight reviews have a higher average score than users that provide more than eight. This can be observed by comparing the user's results (green over yellow) at each bin time of the item progression. Furthermore, by observing the distribution of those users that provided less than eight reviews overtime (green box), we can see that their average score declines over time.

These results suggest that the initial drop in ratings seen on the Amazon platform is driven at least in part by a subpopulation of users who provide few reviews overall and who provide inflated ratings for a product early in its lifetime. We hypothesize that this arises due to the nature of the Amazon rating system's existence in support of product purchases. Figures 3·3a and 3·3b confirm this effect and explanation across Amazon categories.

We note that if this explanation holds, then it should be a consistent property across the Amazon platform. Indeed, we find that this is the case, as shown in Figure 3·3a. All categories from Amazon present an initial drop in ratings and most of them – except for `AZ`-App – have an average rating increase afterward. That figure shows that the initial-decline of ratings is a fairly common feature across categories on the Amazon platform.

We can likewise explore our hypothetical explanation – that a subset of reviewers provide early, inflated ratings – for each of the Amazon categories. The results are shown in Figure 3·3b. The figure shows that the early-reviewer effect is present in every Amazon category, and that it is particularly pronounced in certain product categories (Apps and Kindle books).

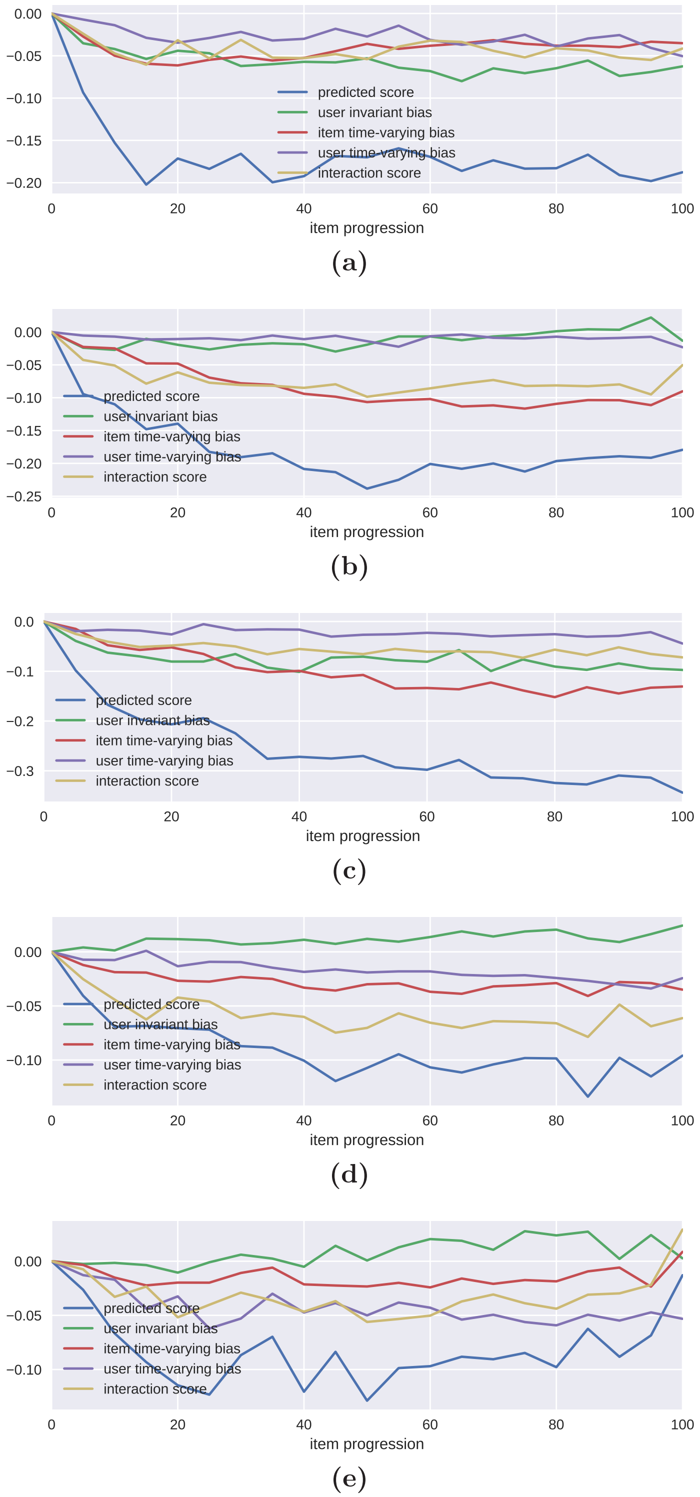
How do recommendations contribute to ratings dynamics? The final factor to consider, as discussed in Section 3.2, is the presence of a recommendation system on a given platform. We expect that if a recommendation system is suggesting items to users, then subsequent ratings for the item should show a higher interaction score because this would reflect an improved 'match' between the preferences of users and the features of the item.

We can assess this effect in two ways: we can ask whether individual items show higher interaction scores over time, and we can ask whether items that have high interaction scores receive more ratings. In the latter case, this may be because more ratings allows the system to do a better job of forming recommendations, and it may be because items that are successfully recommended will garner more ratings.

To ask whether individual items show higher interaction scores over time, we recall from Figure 3·6 and, Figure 3·7 that `RT` and `MT` show essentially zero variation in interaction score (yellow lines). This is consistent with the observation that those platforms are not actively providing users with recommendations that affect which items a user consumes or chooses to rate.

However, that Figure 3·5 shows an interaction score effect for the `AZ` platform. To augment that result, we perform `timeSVD--` decomposition of each of the other Amazon categories. The results are shown in Figure 3·10. Interestingly, it appears that individual items do not show an increased interaction score over time (yellow lines). In general, interaction scores decline somewhat over time.

However, the effect on an individual item may be subtle over time. A more likely effect of a recommender system would occur between the number of ratings an item receives and the interaction score of the item. For this analysis, we return to looking at all items in the dataset (not just those having 100 or more ratings). The results (looking only at interaction score) are shown in Figure 3·11. This figure shows that on the Amazon platform, there is a strong positive correlation between the interaction score (a measure of the effectiveness of the recommendation system) and the number of ratings that an item receives.



(a)

(b)

(c)

(d)

(e)

Figure 3·10: Relative TimeSVD- Components Progression per Category. Subfigures are: (a) Apps; (b) CDs; (c) Kindle; (d) Eletronics; (e) Home .

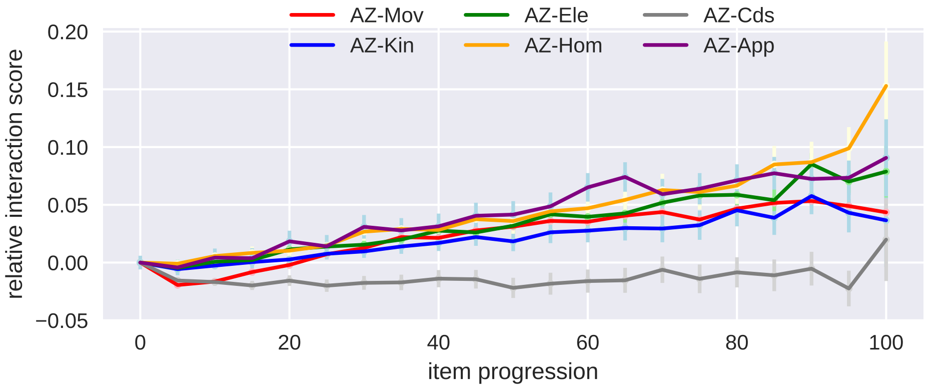


Figure 3·11: Amazon Interaction Score per Category -all items

3.5 Summary of the chapter

In this chapter we've taken a broad look at the factors that drive changes in item ratings in online review systems.

Our results take two parts. First, we characterize the range of ratings dynamics and show how platforms differ. Importantly, different platforms have different and distinctive dynamics. These are preserved when looking at the same sets of items across platforms, and they are preserved when looking across different types of items on the same platform.

Next, we use our model to unravel the factors affecting rating dynamics. First and foremost, we show that changes in user populations are a significant driver of ratings dynamics. In general we observe a trend for user population shifts to increase ratings over time and our `RT` analysis suggests that an important factor is the shift from critics to general users over time. Next, we show that there is in general significant variation in the perceived quality of items over time. This suggests a general trend that may be due to presence of accessible, well characterized reviews (eg in `RT`) or the lack thereof (in `MT`). Then, we show that in the case where ratings are in support of an e-commerce platform (ie, `AZ`) there is a significant tendency for a subset of users who provide few reviews overall to provide early, inflated ratings for items. This is consistent across categories of Amazon products but does not occur in ratings-only sites like Rotten Tomatoes and Twitter. Finally, we find that the presence of a recommendation system on a site like Amazon helps explain the tendency for items (across all categories) that show higher interaction scores to acquire more ratings overall.

Taken as a whole, we show both the complexity behind the dynamics of online reviews and a set of understandable factors that interact to generate that complexity. Hence, we believe that these results provide a framework for interpreting item reviews and how they may be expected to change over time.

Chapter 4

Closed-Loop Opinion Formation

4.1 Introduction

Recommender systems are an increasingly prevalent part of online services, and increasingly mediate access to online resources. Recommender systems are found in systems for online-shopping, video streaming, news feeds, search queries and social media. Recommender systems are employed not just to explicitly give recommendations, but to implicitly guide users, as in the selection and ordering of items in a Facebook news feed.

The term *filter bubble* refers to a narrowed access of information caused by personalization, often in combination with search engines (Pariser, 2011). The term and associated literature raises the concern that recommender systems may have an effect on society, for example by influencing user opinions. Given the prevalence of recommender systems, it is natural to ask whether they can have an effect on user opinions, and what the nature of that effect is.

In this chapter we take this question a step further, and ask how user opinions *and* recommender systems together change as they interact over time. This question arises because many recommender systems are adaptive, making and incorporating observations of users preferences and choices, even as the users are themselves reacting to the recommended items (again, take as an example a Facebook news feed). This constitutes a kind of closed loop, or dynamical system. The central question we ask is: how do user opinions evolve when users and recommender systems each take input from the other, over time?

To answer this question requires significantly abstracting both recommender systems and users in a manner that captures essential properties. Taking recommender systems first, these are usually classified according to how the recommendations are made (Adomavicius and Tuzhilin, 2005): *content-based filtering* systems recommend items based on features of the previous items evaluated by the user, while *collaborative filtering* systems recommend items that people with similar tastes and preferences had evaluated before. Since the latter are based on the opinions of users, our focus is on collaborative filtering systems. We use two different models of collaborative filtering to compare results and explore differences: link based systems and ratings based systems. Link based systems are exemplified by online retailers, which rely on the past records of purchases by the set of all users. Ratings based systems are exemplified by movie, music, and hotel sites sites that recommend new (or previously seen) experiences for a user based on their explicit feedback. These two types of systems are often addressed using the *neighbor* approach and *latent factor* models, respectively (Ricci et al., 2010).

Turning to user modeling, the central effect to be captured is how user opinion about an item is affected by the fact the item has been presented to the user by the recommender system. Here, we study a range of options, from the case where the user's reaction is random, to cases in which users have maximally negative or maximally positive reactions to the item presented. These reactions are captured by the recommender system and used to update its knowledge base.

While there is little work to date that has addressed this question in the form that we pose it, parts of our study have connection to the work in (Dandekar et al., 2013). That paper constructed abstractions of three link based recommender systems, and analyzed the effects on user opinion under certain assumptions. Part of our work looks more deeply at the same systems, examining the effect of key assumptions, and shows effects that differ considerably from what is predicted in (Dandekar et al., 2013). However our work also goes beyond link based systems to ratings based systems, where we find that the set of phenomena (system metrics) to be studied are more diverse.

Beyond abstracting the key system elements, we also must identify the properties of user opinions that are of interest. We study three key properties of the evolving set of user opinions: the *intensity* of individual user's opinions, the *simplicity* of individual user's opinions, and the *divergence* of the opinions of the entire set of users. We make these concrete in the form of specific metrics.

Our results show a surprisingly subtle interaction between properties of recommender system algorithms, actions of users, and initial system state. We show that small differences in algorithms (e.g., whether an algorithm returns the "best" item or merely a "good" item) have a strong effect on whether user opinions undergo simplification as they evolve. We show that the initial distribution of opinions can determine whether users become simplified over time. And we show that recommender systems can cause user opinions to diverge and simplify, but generally only if the recommender system accurately predicts user preferences. On the other hand, we show that a recommender system does not necessarily increase the intensity of user opinions, as long as the system makes good recommendations to the user. We conclude that there is striking richness of interactions that are observed when user opinions and recommender systems form a dynamical system, despite the high level of abstraction necessary for a study like ours. Furthermore, the nature of the effects observed suggests that the increasing prevalence of recommender systems deserves attention and care.

4.2 Framework

We formalize the problem as follows. We consider a system of n *users* and m *items.* Items are objects over which the user has an opinion or has a connection with, e.g., products, movies, books, or news articles.

The knowledge used by the recommender system is held in an $n \times m$ matrix $\mathbf{M}$. In link-based systems, entries in $\mathbf{M}$ are in $\{0, 1\}$, while in ratings based systems entries in $\mathbf{M}$ are in $\mathbb{R}$. In ratings systems, $\mathbf{M}$ is only partially known. In that case, let Ω denote the index set $\{(i_1, j_1), (i_2, j_2), ...\}$ of known (observed) entries of $\mathbf{M}$ and $\bar{\Omega}$ the index set of unknown entries of $\mathbf{M}$. We use $\mathbf{M}^\Omega$ to denote the known entries of $\mathbf{M}$.

In a ratings-based system, a matrix completion algorithm decomposes $\mathbf{M}$ into matrices $\mathbf{X}$ and $\mathbf{Y}$ (k by n and k by m respectively) for a given latent space dimension k such that $\mathbf{C} = \mathbf{X}^T\mathbf{Y}$ and $||\mathbf{M}^\Omega - \mathbf{C}^\Omega||_2$ is minimized.

We model a user's set of preferences or opinions as an s dimensional vector $\mathbf{u_i} \in \mathbb{R}^s$. We think of this vector as the location of the user in a "preference space." This abstract representation can capture a variety of user characterizations. For instance, in a link based system, (such as specified in (Dandekar et al., 2013)) the user vector can be interpreted as a distribution over categories of items to which the user is linked. On the other hand, for a ratings based system using matrix completion $\mathbf{u_i}$ can represent the completed vector of items opinions, i.e., $\mathbf{u_i} = \mathbf{c_i}$ (i-th row of $\mathbf{C}$) or it can represent the projection of the user's ratings in the latent space, $\mathbf{u_i} = \mathbf{x_i}$ (i-th row of $\mathbf{X}$).

Dynamical System The process we study in this chapter is a dynamical system. Hence we introduce time indexing, where $\mathbf{M_t}$ and $\mathbf{u_{it}}$ denote respectively the state of $\mathbf{M}$ and $\mathbf{u_i}$ at time t. The conceptual approach we take is described abstractly as follows.

Let $f(i, \mathbf{M})$ represent a recommender system algorithm; f returns the next item r suggested for user i. Let $g(i, r)$ represent the response of the user to the recommendation; this can be either a rating for the item r or the decision to link to item r (e.g., to purchase item r). This action then provides additional knowledge for the recommender system. Hence after a user has responded to an item, $\mathbf{M}_{t+1}(i, r)$ is set to $g(i, r)$, and Ω is updated to include (i, r).

In this setting we are interested in the dynamical system $\{\mathbf{M}_t\}, t = 0, 1, \ldots$ whose dynamics are governed by

$$\mathbf{M}_{t+1}(i, r) = g(i, f(i, \mathbf{M}_t)).$$

Denote by $\mathit{Pref}(i, \mathbf{M})$ the function that computes the user preference vector $\mathbf{u_i}$. The principal characterization of the system we use is via $\mathbf{u}_{it} = \mathit{Pref}(i, \mathbf{M}_t)$.

This general framework allows for a wide variety of investigations. To assess the impact of recommender system algorithms, we consider two different f functions. The two f functions return either the top rated, or a randomly chosen, item for that user. We also consider a range of g functions, which reflect the degree to which whether users tend to be favorably, or even unfavorably, influenced by the recommended items.

Metrics Previous opinion formation studies have modeled user opinions as scalar values, leading to a one-dimensional representation of user preferences (Dandekar et al., 2013). Nonetheless, empirical studies show that user opinions are better characterized as occupying a higher-dimensional space, e.g., 20 to 40 dimensions (Bell and Koren, 2007). Hence the set of user preferences can be viewed as cloud of points, which we consider to be centered at the origin.

While one-dimensional views of user preferences lead to the single metric of *polarization* to describe opinion dynamics, the more realistic multidimensional view of user preferences provides the basis for a more diverse set of relevant metrics. We introduce three definitions that generalize polarization in different ways to describe how user preferences may evolve over time.

First, we define **intensity** as a per-user metric that captures the strength of a user's preferences. In our point-cloud view, intensity could be conceived of as the distance of the user point from the origin. Hence, given a user i at point $\mathbf{u_{it}}$ at time t, if at time $t' > t$, we have $||\mathbf{u_{it'}}|| > ||\mathbf{u_{it}}||$, we say the user's intensity has increased.

Second, we define **simplification** as a per-user metric that captures the diversity of items that the user prefers, i.e., the spread of user preferences as a distribution. Hence simplification consists of a reduction in the entropy of the user's vector as a distribution, $H(\mathbf{u_i})$. For a given user i

$$H(\mathbf{u_i}) = -\sum_{j=1}^{s} \left(\frac{\mathbf{u_i}[j]}{\sum_{l=1}^{s} \mathbf{u_i}[l]} \right) \log_s \left(\frac{\mathbf{u_i}[j]}{\sum_{l}^{s} \mathbf{u_i}[l]} \right)$$

Then, for a user i at point $\mathbf{u_{it}}$ at time t, if at time $t' > t$, we have $H(\mathbf{u_{it'}}) < H(\mathbf{u_{it}})$, we say the user's opinions have undergone simplification. We also use the term *diversification* as the opposite of simplification, i.e., an increase the entropy of the user's preference vector.

Finally, we define **divergence** as a property of a set of users that captures the similarities among the users' preference vectors, i.e., the degree two which any two user's preferences are alike. To measure this we use the average correlation coefficient over all pairs of user preference vectors, $\bar{\rho} = \frac{2}{n^2-n} \sum_{i>j} \rho(\mathbf{u_i}, \mathbf{u_j})$ where ρ is the standard correlation coefficient. Then, if we have for a set of users at time $t' > t$ that $\bar{\rho}_{t'} < \bar{\rho}_t$ we say that the user set has increased divergence.

4.3 Link Based Systems

A large class of recommender systems can be abstracted using a graph; we call these *link based* systems. A link based system is modeled as a bipartite graph

$G = (V_1, V_2, E)$, where nodes in V_1 represent users, nodes in V_2 represent items, and E is the set of edges, i.e., connection among those nodes. In such a system the recommender algorithm is a function f that takes as input G and a node $i \in V_1$ and outputs a node $j \in V_2$. The representation of those connections varies with the specificity of each system. Edges can be unweighted to simply represent viewing of a video or purchase of an product such as at (Dandekar et al., 2013), or weighted edges representing scores or ratings such as at (Cooper et al., 2014).

In this section we study three abstractions of link based systems introduced in (Dandekar et al., 2013). Our goal is twofold: first, we use these as a comparison case for our study of ratings based systems in the next section; and second we seek to extend and probe the limits of the analysis performed in (Dandekar et al., 2013).

4.3.1 Polarization: theoretical analysis

In (Dandekar et al., 2013), the authors investigate whether certain link based recommender systems have a polarization effect, i.e., whether the recommender system dynamics result in an increased divergence of opinions.

The authors analyzed three random-walk based recommender algorithms inspired by well-known algorithms from literature: SALSA (Lempel and Moran, 2001) (*SimpleSALSA*), Item-based Collaborative Filtering (Linden et al., 2003) (*SimpleICF*), and Personalized PageRank (Page et al., 1999) (*SimplePPR*) , described in more detail below. We follow their framework, in which items have labels $l \in \{$“RED”, “BLUE”$\}$, and there are an equal number of items of each label. That is, $|V_1| = n$ and $|V_2| = m = 2w$ with w items of each label.

These analyses considered cases in which users respond either with or without what was termed *biased assimilation*. In that context biased assimilation specifically means that the probability that a user i accepts an item recommendation is proportional to the quantity of items that i has of that label. On the other hand, without biased assimilation the probability that a user accepts a given recommendation is label-independent. Specifically, let x_i be the fraction of "RED" items owned by i. A recommender algorithm is polarizing with respect to i if: (1) when $x_i > \frac{1}{2}$ the probability that than a recommended item accepted by user i is "RED" is greater than x_i, and (2) when $x_i < \frac{1}{2}$, the probability that the recommended item accepted by user i is "RED" is less than x_i.

The authors in (Dandekar et al., 2013) conclude through analysis that *SimpleSALSA* and *SimpleICF* are polarizing only if users respond with biased assimilation; in contrast, *SimplePPR* is always polarizing. That analysis includes three assumptions: (1) the number of "RED" and "BLUE" items is equal; (2) the set of items is arbitrarily large (i.e., they study the system properties as $m \to \infty$); and (3) the recommender system may recommend the same item multiple times, even if the user has already linked to it. The third assumption is not explicitly stated but is implicit. One of the goals of this section is to show that analysis of these algorithms are quite sensitive to these assumptions and that as a result, conclusions about polarization are necessarily more nuanced.

4.3.2 Model

We encode the link based model in a binary matrix $\mathbf{M}$ such that $\mathbf{M}(i,j) = 1$ iff there exists an edge $e_{ij} \in E$ that connects $i \in V_1$ to $j \in V_2$. Once that there are equal quantities of items of each label we fixed the item label with respect of its position in $\mathbf{M}$, for instance, the first $\frac{m}{2}$ items have "RED" label and consequently last $\frac{m}{2}$ items have label "BLUE".

The recommender algorithms studied in (Dandekar et al., 2013) were defined in terms of a random walk on G. However, analysis becomes clearer if we express the algorithms in terms of a Markov chain. Let $\mathbf{P_{V_i,V_j}}$ be the transition matrix between elements $i' \in V_i$ to $j' \in V_j$. So, $\mathbf{P_{V_1,V_2}}$ is n by m matrix can be calculated from

$\mathbf{M}$ such that each element $p_{ij} = \frac{m_{ij}}{\sum_j m_{ij}}$ and $\mathbf{P_{V_2,V_1}}$ is m by n matrix such that $p_{ij} = \frac{m_{ji}}{\sum_i m_{ji}}$. Since G is bipartite $\mathbf{P_{V_1,V_1}} = 0$ and $\mathbf{P_{V_2,V_2}} = 0$. Thus:

$$\mathbf{P} = \begin{bmatrix} \mathbf{0} & \mathbf{P_{V_1,V_2}} \\ \mathbf{P_{V_2,V_1}} & \mathbf{0} \end{bmatrix}$$

We can now rewrite the random-walk based recommender algorithms as Markov chain based algorithms as follows.

Algorithm 1 *SimpleSALSA* using transition matrix **P**

```
Require: $\mathbf{M}$ and a user $i$ .
1: Compute $\mathbf{P^3}$, a three-step transition on the transition matrix $\mathbf{P}$
2: Choose an item $j$ according to the distribution $\mathbf{P^3}(i, \cdot)$.
3: Return $j$
```

Algorithm 2 *SimpleICF* using transition matrix **P**

```
Require: $\mathbf{M}$ and a user $i$.
1: Compute $\mathbf{P^2}$, a two-step transition on the transition matrix $\mathbf{P}$
2: Choose $k$ according to the distribution $\mathbf{P}(i, \cdot)$
3: Compute $j = \arg\max_s \mathbf{P^2}(k, s)$
4: Return $j$
```

Algorithm 3 *SimplePPR* using transition matrix **P**

```
Require: $\mathbf{M}$ and a user $i$.
1: Compute $\mathbf{P^3}$, a three-step transition on the transition matrix $\mathbf{P}$
2: Compute $j = \arg\max_k \mathbf{P^3}(i, k)$
3: Return $j$
```

Note that none of these algorithms are prevented from recommending an item to which the user has already linked. While some systems such as music recommenders may suggest items multiple times, for many other systems it is not desirable to recommend items that have already been purchased or viewed (e.g., books, movies, news articles).

To study the case in which repeated recommendation of an already-linked item is not allowed, we adjusted the above algorithms during simulation. In each case we simply ensured that the algorithm did not return an item that had been already linked by the user. Denoting Ω as the set of items already linked, we ensure that each item k such that $(i, k) \in \Omega$ is removed from the distribution used in *SimpleSALSA* and eliminated from consideration in the $\arg\max$ computations in *SimpleICF* or *SimplePPR*. In such cases, to ensure that a recommendation is always possible, we replace the zeros in $\mathbf{P}$ with a small value $0 < \epsilon \ll 1$.

4.3.3 Simulation and Analysis

We would like to understand the role of each factor – recommender algorithm, user behavior, and initial system settings – in the formation of user opinion in the linked model G.

For this, we use simulation consisting of the following steps: (i) First, initialize $\mathbf{M}$ according to a probability distribution $q(\cdot)$; (ii) Provide recommendations for all users according to Algorithm 1, Algorithm 2 or Algorithm 3; (iii) For each user, accept the suggested recommendation with some probability p, where p can be a fixed probability (when users respond without biased assimilation), or can vary with the fraction of items with same label that the user has (for biased assimilation). (iv) The matrix $\mathbf{M}$ is updated synchronously with the accepted items . The system evolves (repeating steps (ii) to (iv)) for $T = 1000$ timesteps. We repeat each simulation 30 times and report confidence intervals.

For simulations in this section we define user preference $\mathbf{u_i}$ as a vector consisting of the distribution c_{li} of items the user i is linked to over all labels l. We use two metrics: (i) average user entropy, i.e. $E[H(\mathbf{u_i})] = \frac{1}{n}\sum_i H(\mathbf{u_i})$ where $H(\mathbf{u_i}) = -\sum_{l=1}^{s} c_{li} \log_s(c_{li})$ (ii) the system entropy $H(\mathbf{U})$, where

$$H(\mathbf{U}) = -\sum_{l=1}^{s} \left(\frac{\sum_i c_{li}}{n}\right) \log_s \left(\frac{\sum_i c_{li}}{n}\right).$$

Hence the property we study is simplification (as defined above) which in this case measures the tendency for users to link primarily to one label type versus the other. We use simplification rather than polarization because it measures the degree to which user opinions become more extreme in cases where there may be more than two labels, and hence is more general than polarization.

Case Study 1: including already-linked items

Our first set of results studies the case in which the system is allowed to recommend an already-linked item. The settings simulated were 4700 users and 3700 items. Each user was initialized with an average of 40 known items uniformly distributed between labels. Those settings are inspired in by **MovieLens** (GroupLens, 2015) dataset. The label-independent acceptance probability for non-biased user responses was $p = 1.0$, i.e., the user always accepts the recommended item.

Figure 4·1a and Figure 4·2a show the simplification effects for Case Study 1 over 1000 steps. The figure shows that none of the 3 algorithms show significant variation over time, with or without biased assimilation. This is surprising because the theoretical analysis of (Dandekar et al., 2013) suggests that *SimpleSALSA* and *SimpleICF* should have distinct behavior (not polarizing and polarizing) when comparing label-independent and biased assimilation.

However, a closer look at those simulation outcomes revealed that they were mostly outputting the same recommendation over time.

A simple inspection of the *SimpleICF* and *SimplePPR* algorithms and their respective random process reveals that without constraints preventing the repetition of the same item, and given a finite number of users and items, the selection step (through the $\arg\max$ computation at line 2) of *SimplePPR* and *SimpleICF* is likely to always return the same item; consequently few or no updates are made on $\mathbf{M}$, explaining the constant entropy measured. In contrast, when an infinite number of items and users is considered a higher randomness in the selection step is expected. Therefore,the chances of return the same item is reduced.

This leads to two conclusions: (1) the simplification (or polarization) effects of a real (finite) recommender system are different than those of a idealized infinite system; and (2) the previous conclusions that *SimpleSALSA* and *SimpleICF* can be non-polarizing in some cases need to be re-examined for realistic systems.

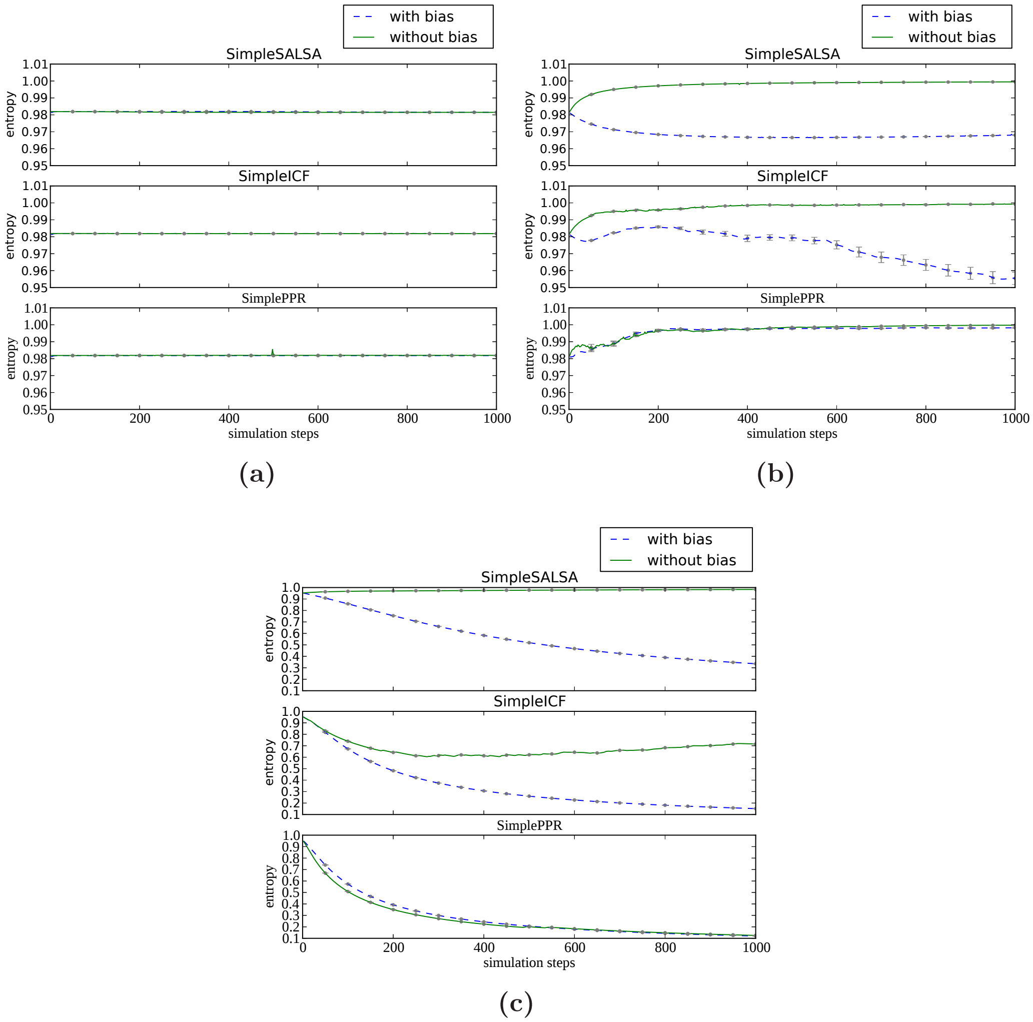
Case Study 2: no already-linked items

Next, in Case Study 2 we look at how results change when the system is not allowed to repeat already-linked items. In all other respects, simulation settings are the same as Case 1.

The simplification of user preference in Case Study 2 is shown in Figure 4·1b. The figure shows that (i) without biased assimilation all algorithms have a diversification effect; (ii) with biased assimilation *SimpleSALSA* lead to simplification, *SimpleICF* has an initial diversification followed by a simplification effect and, *SimplePPR* leads to a diversification effect. This shows that when linked items are not repeated, results are very different from those in Case Study 1, and from the systems analyzed in (Dandekar et al., 2013).

We also analyze for Case Study 2 the simplification of the system, i.e. the diversity of the combined preferences of all users as it evolves over time. Figure 4·2b shows that for*SimpleSALSA* and *SimplePPR* the system has a similar and almost constant effect regardless of whether users respond in a label independent way or with biased assimilation. However, *SimpleICF* suffers a simplification effect when there is biased assimilation.

The diversification effects for biased assimilation in Figure 4·1b for the initial steps of *SimpleICF* and *SimplePPR* may seem counter-intuitive, but they can be understood as resulting from the less-diverse set of recommendations that those algorithms

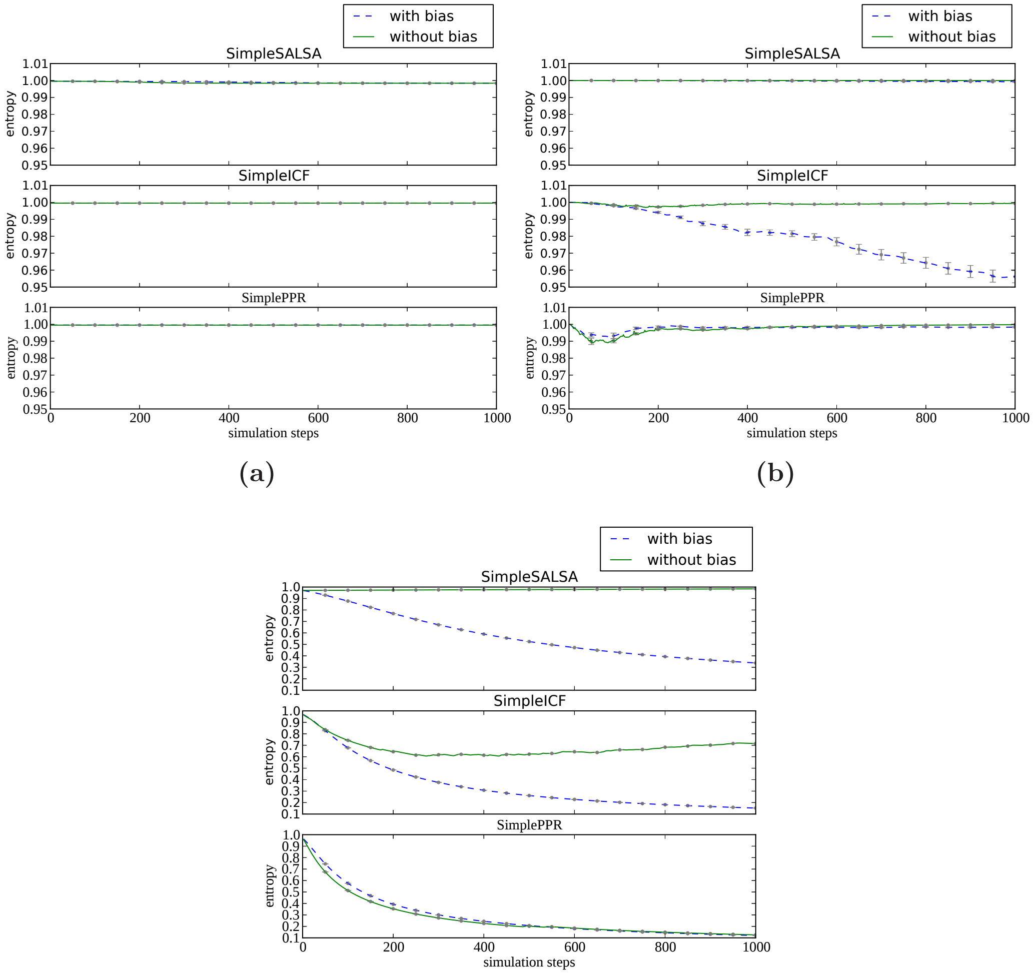


(a)

(b)

(c)

Figure 4·1: Average Entropy: User's Preference. Subfigures are: (a) Case Study 1; (b) Case Study 2; (c) Case Study 4.



(a)

(b)

(c)

Figure 4·2: Average Entropy: System' Preference. Subfigures are: (a) Case Study 1; (b) Case Study 2; (c) Case Study 4.

provide to the user. These algorithm provide less diverse recommendations as a result of the $\arg\max$ step that each employs. We note that the most well connected items are those most likely to be recommended because of their higher values in $\mathbf{P^2}$ or $\mathbf{P^3}$. Although initially there is on average the same quantity of items of each label there are also well connected items of both labels due to randomness in the connection pattern. As a result, the lower randomness of recommendations in *SimpleICF* and *SimplePPR* often results in the system recommending the same item repeatedly until the user accepts it. This leads initially to each user linking to items with both labels approximately equally, regardless of whether the user's assimilation is biased or unbiased. However, after some number of simulation steps (approximately 200 steps) the increase of density leads to an increase of the randomness output from *SimpleICF*. I that case, biased assimilation is able to cause a user to acquire more items of one label than another, resulting in simplification of user preference. Furthermore the greater randomness of *SimpleSALSA* causes a similar biased consumption effect resulting in simplification of individual users' opinions. This is further confirmed by the fact that the overall system sees no change in the label distribution as shown in Figure 4·2b under *SimpleSALSA*.

In summary, this section shows that the tendency of user preferences to simplify or diversify depends on a delicate interplay of the diversity of items suggested by the recommender system and the extent to which users exhibit biased assimilation.

Case Study 3: Variations

To test whether our results in Case Study 2 are representative, we vary the simulation parameters from Study Case 2. In particular, we study the effects of varying;

1. The percentage of items to which each user is initially linked. Previous cases used 1.08%, here we also explored 10%;

2. The label-independent probability of accepting an item in the unbiased biased case. Previous results used $p = 1.0$, here we range p in $\{0.6, 0.8\}$.

None of the above variations affected significantly the trends results presented in Study Case 2, regardless of recommender algorithm or biased/unbiased assimilation. Rather than presenting all results in full, we summarize their similarity by presenting statistics in Table 4.1. The table shows the coefficients of linear regression lines fit to each individual algorithm and response variation. Specifically, Table 4.1 presents the slopes (noted as s) and intercept points (noted as i) for algorithms *SimpleSALSA*, *SimpleICF* and *SimplePPR* (noted respectively as $f = \{1, 2, 3\}$). In the Table the percentage of initial items is denoted as $q(\%)$, the number of users is n, the number of items is m, and the user response probability or biased response indicator is denoted as $g(p)$. Across almost all variants in Table 4.1 we observe similar slopes and intercept points.

Case Study 4: non-uniform initialization

The case studies above demonstrate how different algorithms and user responses evolve in a system in which links to items of different labels are initially uniformly-distributed. We undertake Case Study 4 in order to better understand the role played by the initial distribution of item labels. The setting of Case Study 4 is the same as Case Study 2 except for the initialization step. Each user is still linked on average to 40 items but on average 60% of those links are to items with the first label (”RED”) and 40% are to items with the second label. Figure 4·1c and Figure 4·2c shows the simplification effects over time at the user and system level respectively when there is a uneven starting link distribution and no repeat items are allowed. The figures show first of all that the degree of simplification in this case is dramatically larger than in any of the balanced-link cases. This shows the very strong effect that the initial preferences of users have on the evolution of the system. Second, the figures show that the high levels of simplification are visible both at the user level and at the system level. Overall, the fraction of links that are to RED items grows dramatically at the system level. Third, with respect to the behavior of various algorithms, we note the following. Without biased assimilation *SimpleSALSA* still has a diversification effect on both user's and system level effect measured by the increase of average entropy. However when the user's response is biased the *SimpleSALSA* results in significant simplification at both the user and system level. On the other hand, *SimpleICF* and *SimplePPR* always result in very signficant simplification at the user and system level regardless of the user's response behavior.

The dramatic change of behavior of *SimpleICF* and *SimplePPR* when starting preferences are unbalanced is once more explained by the selection step (through the $\arg\max$ computation at line 2). When there are initially more links to items of one label the most well connected items are also from that label, creating a cycle of same-label recommendations.

4.3.4 Discussion

The failure of the theoretical analysis to consistently match any set of results in our case studies reveals how sensitive the analysis of system dynamics in (Dandekar et al., 2013) is to assumptions and initial state. This can be seen intuitively as well. For example, the operation of *SimpleSALSA* is just to take three steps in a Markov Chain starting from a random state. Intuitively this results in an output vector that interpolates between the initial state and the steady state of the chain. The extreme symmetry of the initial system setup in (Dandekar et al., 2013) suggests that the steady state is the uniform distribution (considered separately over users and items). Creating a new link according to this distribution will move the system to an ever more uniformly connected state – hence the decrease in polarization shown in Figure 4·1a and Figure 4·2a which agrees with the analytical results in (Dandekar et al., 2013). However, the same analysis suggests that the analytical results are strongly dependent on the initial linking pattern, as an imbalance in that pattern will result in an imbalanced steady state, and hence the addition of links in an imbalanced fashion as well.

Our analysis also shows the key role played by diversity of recommendations. Both *SimpleICF* and *SimplePPR* attempt to provide a notion of "best" recommendation through the use of the $\arg\max$ step. On the other hand, *SimpleSALSA* only tries to provide a "good" recommendation (with high probability). While providing the "best" recommendation may seem more optimal in some sense, *SimpleSALSA* shows a much smaller tendency to simplify user preferences, particularly when user preferences start out in an imbalanced state (which seems the likely case in practice).

4.4 Ratings Based Systems

The link based systems considered in the previous section are relatively easy to analyze and interpret. However, in many recommender systems the relation between a user and an item goes beyond a simple binary connection and is expressed in some form of numerical rating. We term such systems as *ratings based systems*. A recommender algorithm in that scenario aims to predict the ratings of the unevaluated items – i.e., anticipate how the user will evaluate the remaining unevaluated items.

One of the main methods used by rating based systems is to provide recommendations using latent factor models (Ricci et al., 2010). Latent factor models estimate the ratings relations between users and items by modeling each in a latent space. The latent vectors are are learned from the data. Ratings are then estimated using inner product of user and item vector in the latent space. This estimation process can be also understood as *matrix completion*.

In a system where user preferences are real-valued, there are a wider array of metrics that are important and can be considered. Our goal in this section is to understand not just how simplification evolves (as in the last section) but also how intensity and diversity evolve in time.

4.4.1 Model

Let $\mathbf{M} \in \mathbb{R}^{n \times m}$ be a ratings matrix of n users over m items, λ_{MIN} be the minimum ratings value, λ_{MAX} be the maximum ratings value, and a the completion algorithm that decomposes $\mathbf{M}$ into the factors $\mathbf{X}$ and $\mathbf{Y}$ where $\mathbf{C} = \mathbf{X^T Y}$.

Once $\mathbf{C}$ is computed, a typical way to provide a recommendation r to a user i is using an algorithm f that recommends the unevaluated item $C(i,j)$ with highest predicted rating. We denote that algorithm as *RecBEST*; it is described below (Algorithm 4).

As a comparative case we define *RecRAN* (Algorithm 5) to be the recommender algorithm f that simply recommends a random unevaluated item.

Algorithm 4 RecBEST using **C**

```
Require: $\mathbf{C}$, $\mathit{\Omega}$ and a user $i$ .
Compute $j = \arg\max_j C[i,j]$ such that $(i,j) \in \bar{\mathit{\Omega}}$.
Return $j$
```

Algorithm 5 RecRAN using **C**

```
Require: $\mathbf{C}$, $\mathit{\Omega}$ and a user $i$ .
Choose randomly $j$ such that $(i,j) \in \bar{\mathit{\Omega}}$.
Return $j$
```

We seek to capture a range of possibilities for the response of a user to a recommendation. These essentially reflect how a user's opinion changes in response to evaluating an item (e.g., viewing a movie or reading a book). We model the user response $g_x(i,r)$ as a probabilistic function of $x \in [0,1]$. The user signals her evaluation by providing either a rating of λ_{MAX} or λ_{MIN}. The parameter x determines how often a user tends to be positive about a recommended item. That is, for a given user i and a recommendation r, $\mathbf{M}$ will be updated accordingly to $g_x(i,r)$, i.e., $M(i,r) = g_x(i,r)$, where:

$$g_x(i,r) = \begin{cases} \lambda_{MAX}, & \text{with probability } x \\ \lambda_{MIN}, & \text{with probability } (1-x) \end{cases} \tag{4.1}$$

4.4.2 Datasets

The previous section showed that system dynamics can be strongly influenced by the initial system state. Hence we conclude that it is important to initialize the system in a realistic manner. As a result we use previously captured datasets as the system initialization in this section. We use three datasets that include ratings relations between users and items to define the starting point matrix $\mathbf{M}$ and initial Ω.

The **MovieLens** dataset is collected from a non-commercial web movie recommender (GroupLens, 2015). We selected a relatively dense subset of this dataset, consisting of users that have at least 40 ratings, which we denote as `dataML`. In total `dataML` has 4736 users, 3706 movies, and 962682 ratings. Additionally we scaled the ratings to center them at zero, changing the initial range from [1:5] to [-2:2].

The **MovieTweetings** dataset is collected from well-structured movie evaluation tweets on Twitter from 2013 until 2015 (Dooms et al., 2013). We selected a relatively dense subset of this dataset of movies that have at least 10 ratings and users that have at least 40 ratings, which we denote `dataMT`. This dataset has 2604 users, 3703 movies and 218302 ratings. We centered the ratings in `dataMT` by rescaling them from [1:10] to [-5:5].

The **BookCrossing** was collected in a 4-week crawl from the Book-Crossing community (Ziegler et al., 2005). We used a relatively subset of this data excluding null inferred ratings and selecting user with at least 40 ratings over books with at least 5 ratings, which we denote `dataBX`. This dataset has 294 users, 2764 books, and 20040 ratings rescaled from [1:10] to [-5:5].

4.4.3 Simulations

We again study the closed-loop dynamics between recommender system and users in simulation.

In our simulations all matrix completion procedures were performed using LMaFit (Wen et al., 2012) with a latent space $k = 20$. We define the user preference vector as the completed vector of item opinions ($\mathbf{u_i} = \mathbf{c_i}$) and observe how $\mathbf{u_i}$ evolves over time through an individual dynamic simulation, i.e., just one user evolves in time while other user ratings are not changed. We observe how the user preference evolves in the dynamical system at an individual level (through intensity and simplification) as well as at a system level (through divergence).

We analyzed the recommender algorithms $f \in \{RecBEST, RecRAN\}$. over the probabilistic users response $g_x(i, r)$ for $x \in \{0.0, 0.1, \ldots, 1.0\}$. Using those settings we evolved each user for $T = 400$ iterations where at each step a new recommendation was made and evaluated by the user according to rating policy g_x. Furthermore some individual and collective metrics were computed from $\mathbf{u_i}$ at each step. This framework for our analysis is described in Algorithm 6.

Intensity

First we consider how the intensity of the user preference vector varies over time and how the choice of user response and recommender system influence in those changes.

Figure 4·3 captures the intensity measures computed for all of our simulations. Each square from each plot of Figure 4·3 represents one metric observation, while the color of the square indicates the metric value as indicated in the color map. Furthermore, the x-axis in each plot represents the simulation step (ranging from 1 to 400) and, the y-axis in each plot represents the probability with which the user

Algorithm 6 Ratings based dynamical system

```
Require: A partially-observed matrix of ratings $\mathbf{M}^{\Omega}$; a recommender system algorithm $f$; a user response function $g_x$; a preference space size $k$; and a number of iterations $T$.
for $i \in$ $Users$ do
    $\mathbf{R} = \mathbf{M}$
    for (step=1:T) do
        $(\mathbf{X}, \mathbf{Y})$ = LMaFit($\mathbf{R}, k, \Omega$)
        $\mathbf{C} = \mathbf{X}^{\mathbf{T}}\mathbf{Y}$
        $j = f(\mathbf{C}, i, \Omega)$
        $\mathbf{R}(i, j) = g_x(i, j)$
        $\Omega = \Omega \cup (i, j)$
        Output Intensity and Simplification of $\mathbf{c_i}$
        $S$(step, i) = $\mathbf{c_i}$
    end for
end for
for (step=1:T) do
    Output Divergence of $S$(step)
end for
```

response g_x was set. Thus each horizontal set of points represent one particular simulation. Each plot of Figure 4·3 represents a set of simulations results from a given algorithm and dataset. Different set of plots rows correspond to the dataset used to initialize M – respectively `dataML` , `dataMT` or `dataBX` . The first two columns of plots group simulation results from *RecBEST* and *RecRAN* respectively. The last column shows the pointwise subtraction of *RecRAN* from *RecBEST* for comparative purposes.

The reason for the last column of plots is that the *RecRAN* simulations measure the effects on user opinion when randomly chosen items are evaluated by the user according to the rating policy g_x. Thus, this case can be considered to capture the effect of a user viewing and rating items without the influence of a recommender system, but rather in completely random fashion. Therefore the subtraction of the simulations results (*RecBEST* - *RecRAN*) – noted in this work as *BEST-RAN* – is a

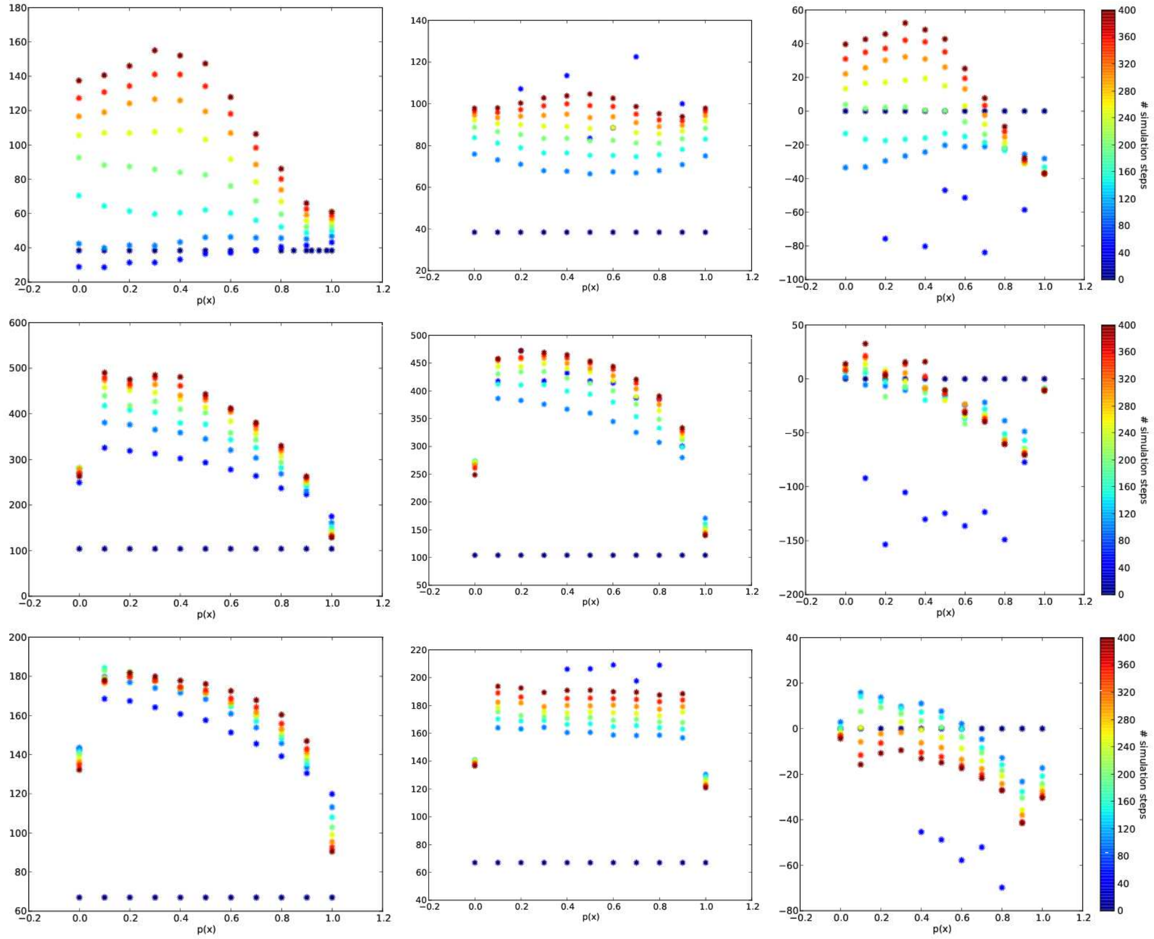


Figure 4·3: Average Norm of User's Preference

measure of influence of the recommender system *algorithm* on user opinion.

All the intensity observations represented at Figure 4·3 were computed by averaging the norm of the users preference vectors for a given time step.

The figure show a number of results. First, for both algorithms (*RecBEST* and *RecRAN*) and for all datasets, regardless of the user's response, we can observe a general increase in intensity over time – marked by the vertical increase of average norm value when the time color indicators also increases.

Second, for *RecBEST* (first column) the increase in intensity of opinion is maximized for probabilities x intermediate between 0 and 1. The *RecBEST* system increases intensity the least when $x = 1$, the user always agrees with the system and rates the items presented highly.

These two effects together mean that for *BEST-RAN* (third column) there is a generally decreasing relationship between the user opinion intensity and the degree to which the user agrees with the recommender algorithm. In other words, when the user agrees with the recommendations made by the system, the user's opinion intensity increases, but the increase is lower than if there had not been a recommender system in the loop. On the other hand, when the user disagrees with the recommender system, the user's intensity of opinions increases more than if there had been no recommender system in the loop. We conclude from this that a recommender system does not necessarily increase the *intensity* of user opinions, as compared to random recommendations, as long as the system makes good recommendations to the user.

Simplification

Our second analysis concerns how the user's preference vector behaves over time as a distribution, which is captured as simplification (or diversification) of opinions. Figure 4·4 compiles the results from the simplification measures of the user preference vector. The figure uses the same representation for squares, colors, plots and plot positions as Figure 4·3.

The figure shows a number of effects in opinion dynamics. First of all, for both algorithms (*RecBEST* and *RecRAN*) and for all datasets regardless the user response, we can observe a general increase of entropy over time. In another words, user opinions get more diverse under the influence of either a recommender system or a random presentation of items.

Second, the influence of dataset (i.e., system initialization) has a much stronger effect on diversity (as compared to, eg, intensity). The changes in diversity are highly varied across datasets (although diversity always increases).

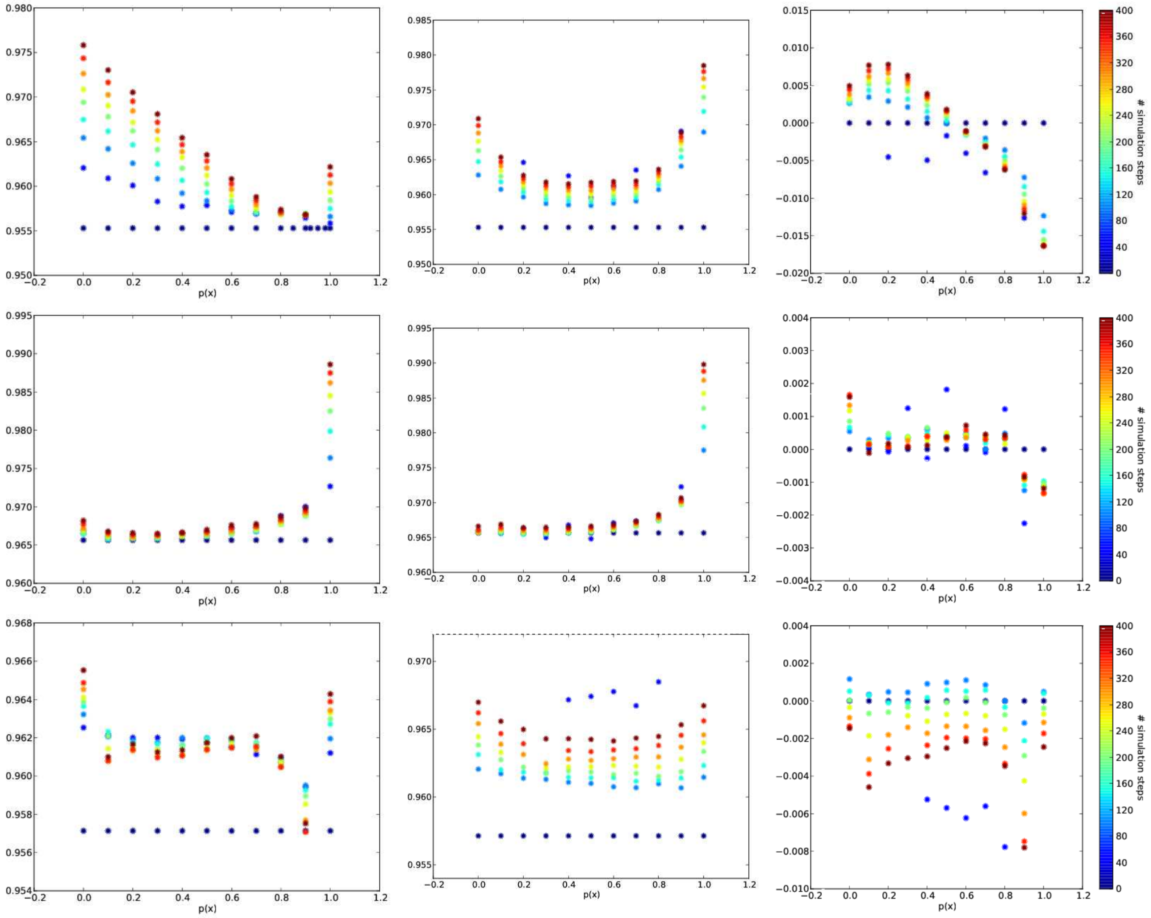


Figure 4·4: Average Entropy of User's Preference

However, the third column (*BEST-RAN*) shows that the relative influence of the recommender algorithm *RecBEST* over users for `dataML` and `dataMT` is that when y increases, the average entropy decreases. This means that, when the user agrees more with the recommendations its opinions become simpler, as compared to a random presentation of items. Even for `dataBX`, where there is no strong correlation between the influence of the recommender system and the user response g_x, user opinion always becomes simpler under the influence of the recommender system than under a random item presentation.

These results emphasize the importance of comparing the effect of a recommender system to an alternative. While a dynamical system involving a recommender system tends to result in diversified user opinions, the same is true of a dynamical system that does not involve a recommender system. Only by comparing the two do we see how the recommender system decreases the diversity of user opinion.

Divergence

Our last analysis is with regard to how closed loop dynamics shapes the user preference vectors as a set. We want to understand the conditions under which the user preference vectors become more similar to the others, i.e. we want to observe when there is loss or increase of individuality from the user preference vectors set.

Figure 4·5 presents the average correlation of user preference vectors, with plots using the same representation for squares, colors, plots and plot positions as in previous figures.

We note that for both algorithms (*RecBEST* and *RecRAN*) and for all datasets, the average correlation coefficient decreases for intermediate values of x, while it increases when x is close to either 0 or 1. Thus, unless the user does not fully agree or disagree with most recommendations, the set of user opinions tends to diverge. However, the influence of the recommender system as compared to random recom-

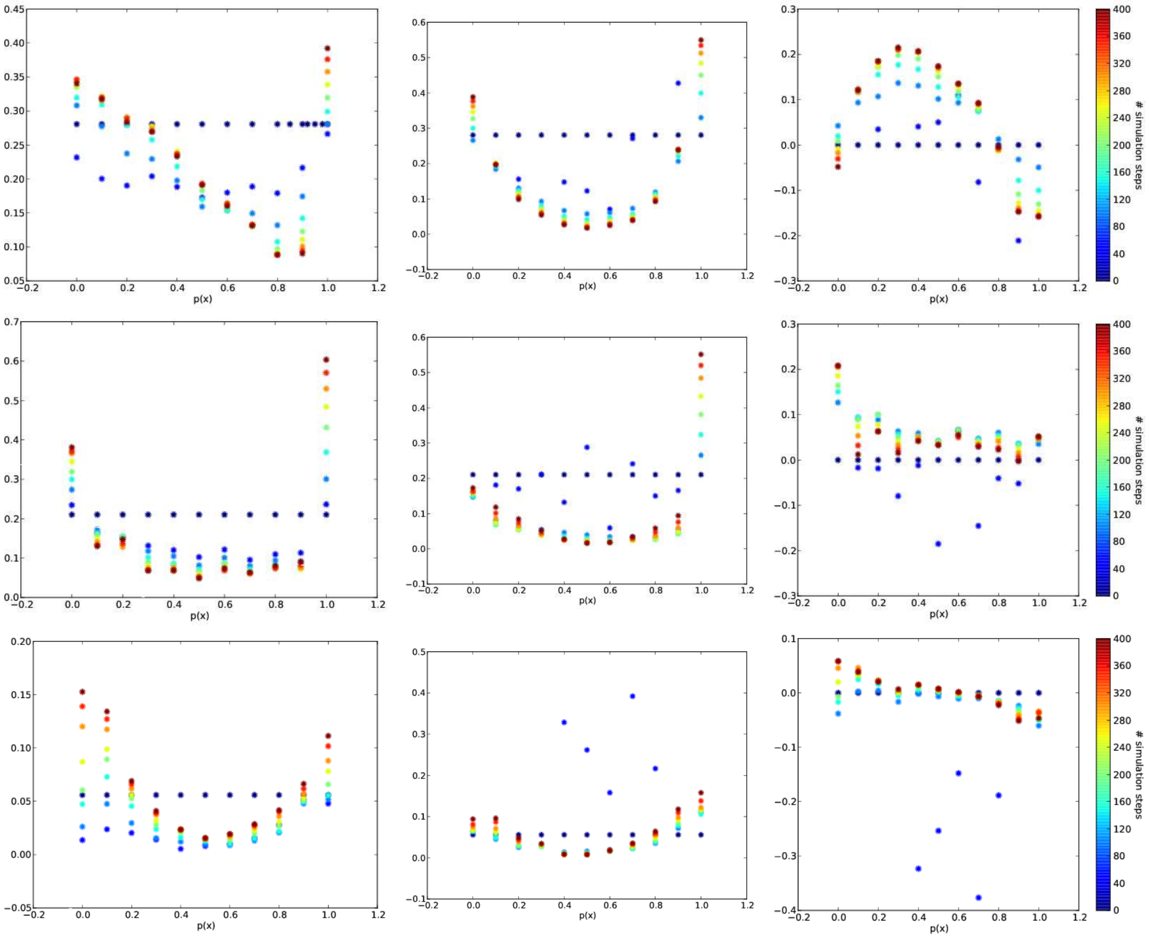


Figure 4·5: Average Correlation Coefficient of User's Preference mendations, as shown in the third column (*BEST-RAN*) is quite different. That column shows that as the user's agreement with recommendations increases, there is a consistent increase in opinion divergence. This effect is strong enough that opinions always increase in divergence, compared to random item presentations, when the user fully agrees with the recommendations made by the recommender system.

4.4.4 Discussion

The conclusions from this section and the previous section together show consistencies that reinforce a number of high level conclusions. The principal conclusion concerns the tendency of recommender systems to simplify user opinions. This tendency is observed generally over both link based and ratings based systems, although it comes with some caveats. First, recommender systems that recommend the "best" item at any given time (eg, *SimpleICF*, *SimplePPR*, and *RecBEST*) have a much stronger simplifying effect than systems that return "good" recommendations according to some probability (*SimpleSALSA*). Furthermore, the simplifying effect is strongest when the recommender system does a good job of predicting user preference (e.g., when $x = 1$ in *RecBEST*).

Second, the initial state of the system when the recommender system starts has a strong effect on the eventual outcome. In the case of *RecBEST*, opinion diversity is strongly affected by the dataset used, and in the cases of *SimpleICF* and *SimplePPR* the simplification effect is highly pronounced when the system does not state in a perfectly balanced configuration.

Turning to the other metrics, our results suggest that the ratings based system tends to cause user opinions to collectively diverge and to become individually less diverse, but *only* when the system accurately predicts the user's preferences. If the system is less effective in this regard, it can cause the opposite effect, namely, the to cause user opinions to become more similar to each other, and to become individually more diverse. On the one hand, the conclusion that opinions can diverge and become less diverse is potentially concerning, but the realization that the recommender system must be quite accurate for this to happen raises questions about whether this effect is likely to occur in practice.

4.5 Summary of the chapter

In this chapter we studied the closed-loop dynamics between recommender systems and users across a wide variety of system models and configurations. We proposed three metrics – intensity, simplification and divergence – to capture important properties of user opinions as they evolve in such a system. By studying a wider range of settings, we extend previous work (Dandekar et al., 2013) and show that its conclusions do not always generalize to more typical settings (eg., when already-linked items may not be recommended, or when the system starts with nonuniform user preferences). Further, comparing results for link based and ratings based systems, we identify common features of recommender systems that tend to simplify user opinions. We also show that under certain circumstances, recommender systems can act to cause user opinions to diverge and become less diverse, but this is not always the case, and more study is needed to determine whether in practice the conditions necessary for divergence and simplification of opinions do actually occur.

Our work has a number of limitations that suggest the need for future study; in particular, further theoretical analysis of both link based and ratings based dynamics seems worthwhile and potentially feasible. Nonetheless, our results show that the dynamics of recommender systems over time are complex and varied, and that there is potential for such systems to affect user opinions in subtle ways. Combined with the increasing prevalence of recommender systems, our results suggest that better understanding of closed loop opinion formation is an ongoing and important problem.

Table 4.1: Case Study 3: Comparison over parameter variations.

n	m	$q(\%)$	f	$g(p)$	slope	intercept
4700	3700	1.08	1	0.6	8.4e-06	0.99
4700	3700	1.08	2	0.6	8.4e-06	0.99
4700	3700	1.08	3	0.6	6.8e-06	0.99
4700	3700	1.08	1	0.8	7.5e-06	0.99
4700	3700	1.08	2	0.8	7.8e-06	0.99
4700	3700	1.08	3	0.8	1.3e-05	0.99
4700	3700	1.08	1	1.0	6.6e-06	0.99
4700	3700	1.08	2	1.0	6.7e-06	0.99
4700	3700	1.08	3	1.0	1.1e-05	0.99
4700	3700	1.08	1	bias	-5.3e-06	0.98
4700	3700	1.08	2	bias	-3.2e-05	0.99
4700	3700	1.08	3	bias	7.4e-06	0.99
4700	3700	12	1	1.0	9.5e-07	0.99
4700	3700	12	2	1.0	5.7e-07	0.99
4700	3700	12	3	1.0	1.0e-06	0.99
4700	3700	12	1	bias	-2.1e-07	0.99
4700	3700	12	2	bias	-7.2e-07	0.99
4700	3700	12	3	bias	7.4e-07	0.99

Chapter 5

How YouTube leads users away from reliable information

5.1 Introduction

Currently, much of the information accessed online is mediated by some kind of recommender system. The increasing and widespread use of recommendation systems has raised concern about how possible biases existing in recommendations can impact worldwide information and public opinion formation.

As a result, research has begun to investigate how personalization can impact the nature of information that is accessed by individuals. One concern is the narrowing of information diversity through the creation of 'filter bubbles' (Pariser, 2011), (Bakshy et al., 2015), (Hannak et al., 2013), (Spinelli and Crovella, 2017), (Le et al., 2019). More recently, the increasing proliferation of unreliable information (Soroush Vosoughi, 2018), (Andrew Guess, 2018), especially on social media, has been adding a new dimension to recommender system social impact and has been increasing the importance of understanding recommendation policies used on such platforms.

Social platforms such as Facebook and YouTube optimize their recommendations to maximize engagement, while commercial platforms such as Amazon seek to drive purchases. However, although most recommendation algorithms are designed for value-neutral objectives such as engagement and commerce, the resulting recommendations can potentially promote content that is factually unreliable or socially harmful. In this regard, the popular press has recently exposed odd behavior of the Amazon and YouTube recommender systems, including promoting radical, extreme, or unreliable content (Tufekci, 2018), (Lewis, 2018),(Chaslot, 2018), (Bergen, 2019), (Diresta, 2019) .

YouTube is one of the most significant sources of socially-generated information globally, with over 1.9 billion logged-in visitors each month and more than a billion hours of video watched every day (YouTube, 2019). However, because of YouTube's revenue model, the nature of its recommendation policies is fairly opaque.

In this chapter, we seek to move beyond the anecdotal descriptions in the popular press and study the nature of YouTube recommendations quantitatively. We study YouTube recommendations empirically, focusing on socially-impactful dimensions – particularly, recommendations for reliable versus unreliable information sources. To this end, we design and implement a data collection framework to simulate users watching a sequence of recommended videos on YouTube under various experimental conditions. We then classify the channels from the recommended videos in terms of the reliability of their content. Finally, we analyze the empirical results to quantify the extent to which YouTube recommendations shift users away from reliable towards unreliable and even extreme content.

Recommender systems are successful to the extent that they can employ information about users, allowing recommendations to be personalized. At the same time, many users seek to protect the privacy of their personal information while online. Hence, one of the central issues we explore in this chapter is the tension between privacy and the nature of recommendations. To that end, our experimental conditions vary in the degree of privacy that our simulated users employ.

Our first contribution is to *quantatively* demonstrate how YouTube's recommendations generally “lead away” from reliable information sources, including a tendency to direct users over time toward video channels espousing extreme or unscientific viewpoints. By quantifying this effect, we demonstrate that in most cases YouTube leads users away from reliable information very quickly. That is, most of the change in the reliability of information takes place within the first few recommendations provided by YouTube.

Our second contribution is to measure the effect of user privacy on YouTube recommendations. While many users may consider privacy desirable, we show that protecting privacy has a major drawback: it drastically increases the “leading away” effect of YouTube recommendations. We show that the increase in the proportion of unreliable content increases by a factor of 2× to 3× for users who preserve their privacy while viewing videos. We quantify this effect along various dimensions, including its dynamics in time, and show how pervasive the tradeoff between privacy and unreliability of recommendations is in the YouTube recommendation process.

Finally, we dive into specific questions designed to explore the robustness of these contributions. We examine how the “leading away” effect depends on the specific topic being explored by the user, showing that “leading away” takes place for most of the topics we study, although to varying degrees. We also show that the widely publicized changes made by YouTube to their recommendation policies in January 2019 decreased but did not eliminate the “leading away” effect.

5.2 Methods

5.2.1 Data Collection

As mentioned above, we study YouTube’s recommendation strategies by following *chains* of recommendations made by YouTube. Starting from a specific *search query,* we simulate a user who watches the resulting video and then selects one of YouTube’s recommendations to watch next. Each chain is collected under a specific *privacy scenario*, and the next video to watch in each case is selected from the list of recommendations according to a *video selection* strategy. We explain each of these experimental aspects in the following subsections.

Privacy scenarios

In order to explore how YouTube recommendations change as a function of what user features are visible to YouTube, we consider four privacy scenarios:

Logged. The user identity is exposed by being logged into a Google account. We used a single university-provided Google account.

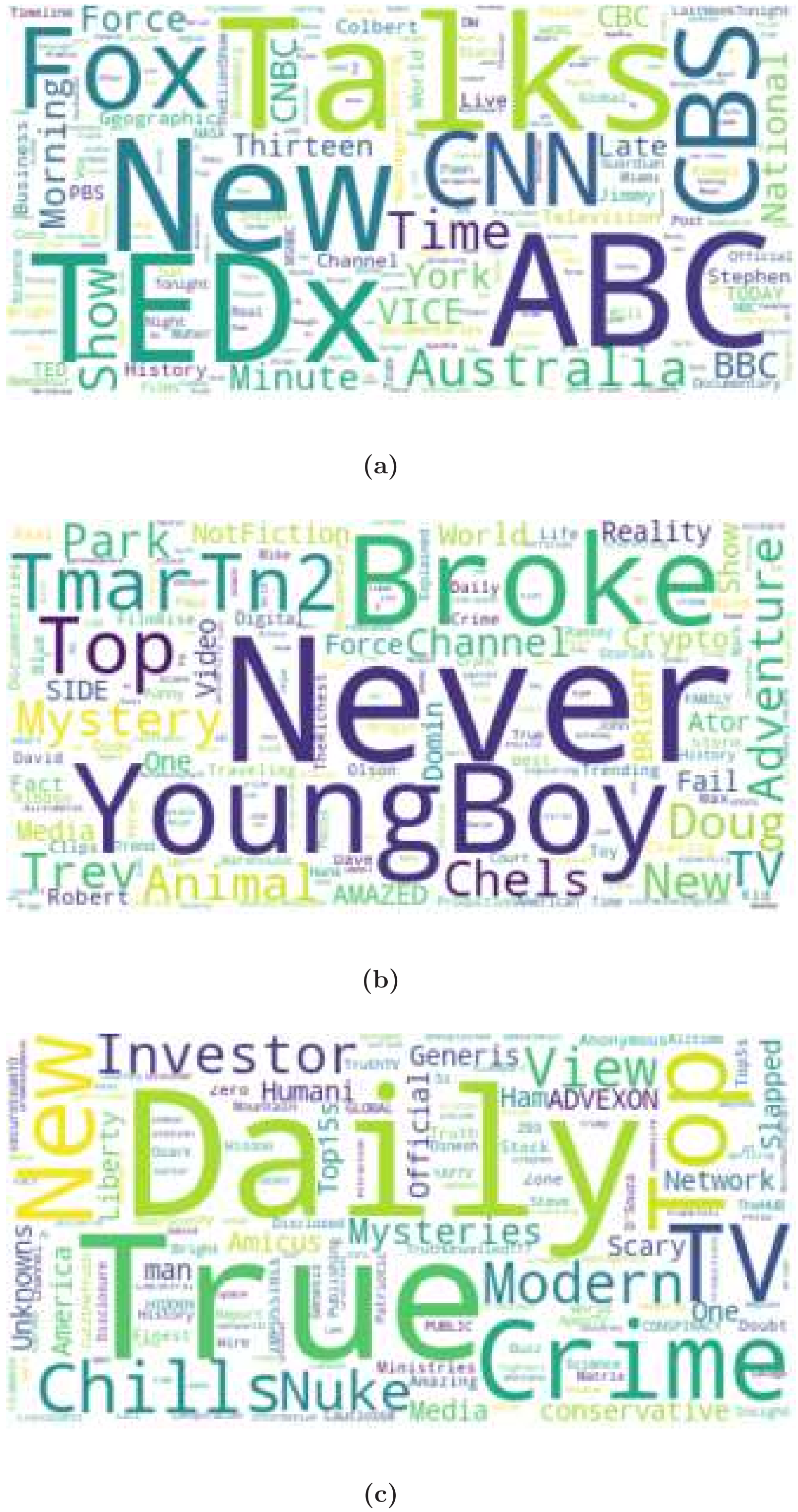
Normal. The user has not logged into a Google account, but uses normal browsing mode which could be potentially tracked by cookies.

Private. The user has not logged into a Google account and uses a private browser session that disables cookie placement.

Tor. The user has not logged into a Google account and uses a private session in a Tor-enabled browser that obfuscates the user's IP address by passing through the Tor network.

Search queries

Each collected video chain starts with a search query. For these queries, we use the top 10 News Google Searches of 2017 in the United States (Google, 2018). We choose these because they represent a set of queries that would be likely as starting points for watching videos on YouTube. The search queries used were: *Hurricane Irma*; *Las Vegas shooting*; *Solar Eclipse*; *Hurricane Harvey*; *Bitcoin Price*; *North Korea*; *Hurricane Jose*; *Hurricane Maria*; *April the Giraffe*; and *DACA*.



(a)

(b)

(c)

Figure 5·1: Channel Name Word Clouds by Classification. Subfigures are: (a) Trustable; (b) Neutral; (c) Extreme.

Video Selection

YouTube's recommendations are provided in a list on the right side of the screen while a video is playing, which we call the *recommedation list.* An important aspect of our experiment is the choice of how the next video to be watched is selected from this list (*video selection.*)

"Auto-play" mode in YouTube simply plays the top item in the recommendation list. This mode is the default behavior of YouTube and is followed when no other user action is made. Hence, we treat the top item as the one most strongly recommended. This video selection strategy is *top item.*

In contrast, to understand the impact of the recommendations rank, we consider a strategy in which the video with the lowest ranking is the one chosen for viewing next. We refer to this video selection strategy as *bottom item.*

We refer to a sequence of videos watched in this way after a single query as a *chain.* In our video selection executions we avoid video repetition inside a chain. Then, if the selected video to play next had already played in the current chain we choose instead the highest unplayed video (or lowest unplayed video) when the video selection is *top item* (*bottom item*).

Features collected

Each YouTube video is published by a YouTube channel. Channels are either based on Google personal accounts or Google brand accounts. Channels with more than 100,000 subscribers that belong to an established creator or are the official channel of a brand, business, or organization can receive a YouTube verification badge checkmark upon request (Google, 2019a).

For each video viewed during our experiments, we collect a set of features. Features are either derived from the video or its channel. The video features we collect are: the current date, the video publication date, the video number of views, the video number of likes, the video number of dislikes, the video number of comments, the video title, the video duration, the video description, the video content category (selected on video publication), the channel identifier, the channel title, the channel number of subscribers and, the channel verification badge status.

Data Collection Process

Putting all the above parts together, the overall structure of our data collection process is given in Algorithm 7. The framework is implemented in python and uses Selenium to simulate user behavior.

Algorithm 7 YouTube Data Collection

```
Require: Privacy scenario, Search term, Selection Strategy
Perform a search query on YouTube and get its recommendation list.
repeat
    Select video from list according to selection strategy.
    if advertisement appears then
        Wait for the end of the ad or skip it.
    end if
    Watch selected video for up to 5 minutes of elapsed time
    collect data about the video and its channel
    get new recommendation list
until video chain reach 20 videos
```

5.2.2 Classification

We classify each recommended video according to its channel. We place each channel into one of three categories: *trustable*, *neutral* or *extreme*. Each channel encountered in our data was classified manually. Manual classification was done via inspection of the most popular movies of the channel as well as the channel description. The criteria we used for channel classification are:

Trustable. Channels identified as trustable are channels from established news sources. Most trustable channels are run by news sources from television, or are credible scientific channels that provide content with externally checkable references.

Extreme. Channels that are identified as extreme are those that have content that deny established scientific knowledge, incite hate or promote fake news.

Neutral. Neutral channels are all other channels – those that are neither trustable neither extreme.

In Figure 5·1 we illustrate of the type of channel in each classification using various word clouds based on channel names. Figure 5·1a shows a word cloud of *trustable* channel names, displaying traditional news sources such as *ABC*, *CBS*, *Fox*, and *CNN*. Figure 5·1b shows a word cloud of *neutral* channel names, and is dominated by entertainment – music and gaming channels – such as *YoungBoy Never Broke* and *TmarTn2*. The word cloud of *extreme* channel names, in Figure 5·1c, shows that extreme video channels use attention-getting names such as “True”, “Mysteries” and “Top”.[1]

5.2.3 YouTube Recommendations

During the period of our study, YouTube generated recommendations using a deep neural network that implements a two-stage approach of candidate generation followed by ranking (Covington et al., 2016). This approach was designed to deliver high performance on key metrics: precision and increased watch time, while handling the challenges of scale, freshness and noise.

YouTube has stated that is continuously working on improving its search results and recommendations using user feedback, external evaluators trained using Google public guidelines (that evaluate the content quality and publisher reputation) and other signals. Relevant to our study, in January of 2019 YouTube announced on its

[1] To allow readers to examine our channel classifications in detail, as well as to reproduce our results, all data and code in the form of Python notebooks will be released upon paper publication.

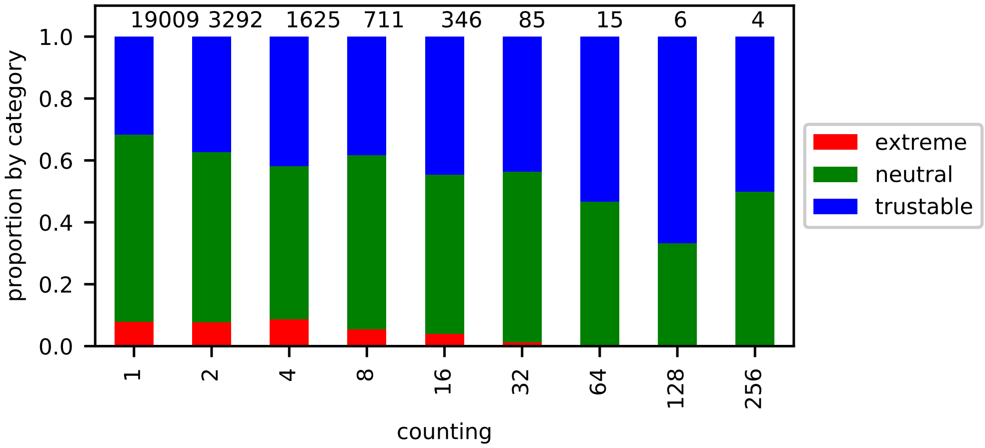


Figure 5·2: Binned counts of the number of appearances of each video, and classification of videos in each bin.

blog that it would reduce recommendation of borderline content and content that could misinform users in harmful ways (Google, 2019b). This fell in the middle of our study period, an event that we analyze in detail in Section 5.4.5.

5.3 Dataset

Our dataset consists of a set YouTube chains collected (as specified in Section 5.2.1) between October 2018 and April 2019. Each experimental setting was replicated 256 times, with one chain collected each time. As a result, the dataset consists of 4 (privacy scenarios) × 2 (selection strategies) × 256 (chains) × 20 (videos per chain) = 40,960 videos. There were 25,091 unique videos in this set. The 10 search queries were evenly distributed across replications.

Although our experiments use a small set of search queries, the videos and channels that are collected are broadly distributed. We show this in Figure 5·2, which bins videos according to how many times each appeared in the dataset. Each bar shows the distribution of classification for the videos in that group, and numbers at the tops of bars show how many videos fall in each group.

We note that most of the videos appear only a few times. For instance, 46.4% of the videos were recommended just once, and only 1.8% of the videos were recommended 16 or more times. The figure also shows that videos belonging to *extreme* channels received relatively few recommendations overall, while videos pertaining to *trustable* channels receive more recommendations – which can be observed by the increasing proportion of this classification for higher-count bins.

5.4 Results

In this section we present the results of our analysis of the effect of YouTube recommendations on the reliability of content seen by users. As described in Section 5.2, our focus is on how reliability of information changes as users follow YouTube's recommendations, and how those effects vary as a function of the privacy scenario.

In our results, we present reliability classification in two formats. First is the unreliability score, computed by assigning each classification *trustable*, *neutral* or *extreme* the number -1, 0, and 1, respectively. We then take the resulting average over time t. The second format is the ternary plot. In the ternary plot, each side of the triangle represents one of the three classification types and each point within the triangle represents a particular proportion across the classifications. We denote this proportion across the classification as the proportion mix. The axis values grow counter-clockwise. For any point, the corresponding fraction values for each classification type can be obtained by the projection of this point into the corresponding axis. Point projection follows a parallel line to the triangle side where this axis has value 0 (clockwise side), for instance, the *neutral* projection line is parallel to the *extreme* triangle side.

5.4.1 YouTube recommendations lead away from trustable sources

Our first result shows that YouTube's recommendations guide users toward less-reliable sources over time. To demonstrate this, in Figure 5·3 we show the characteristics of the sequence of recommended videos, aggregating all privacy scenarios with *top item* video selection. In this Figure, both formats of variation in classification – unreliability score and ternary plots – are displayed. In Figure 5·3a the x-axis denotes the order in the video sequence – going from 1 to 20 – and the y-axis shows the average values of the unreliability score.

The gray line in Figure 5·3a shows that the unreliability score ranges from approximately -0.58 at the beginning of the sequence to -0.23 towards the end of the sequence. This increasing trend shows how dramatically YouTube's recommendations guide users away from reliable channels.

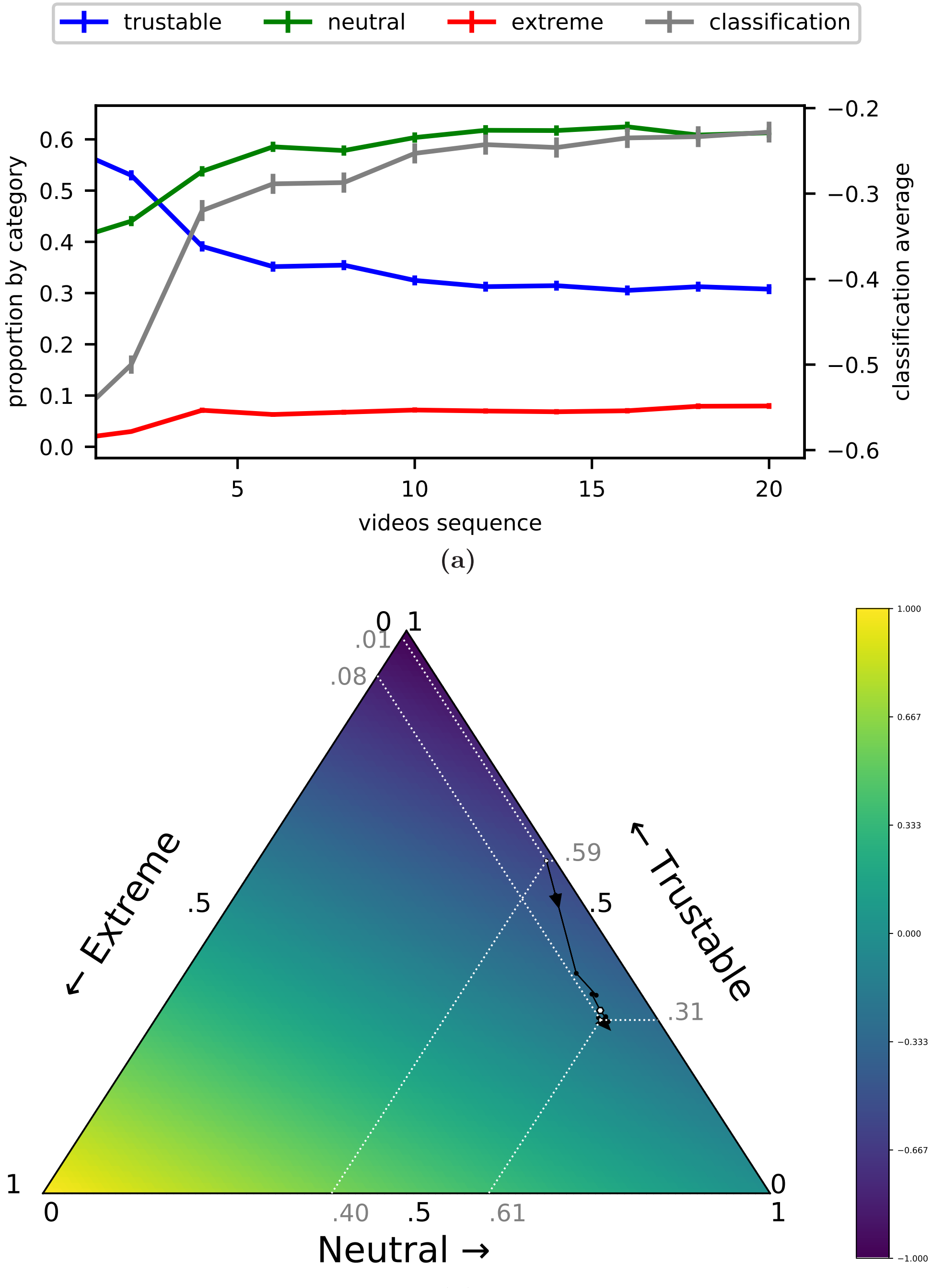
The shift away from reliable videos occurs mainly because trustable channels are mainly replaced with content from neutral channels, along with a slight increase in content from extreme channels. This is illustrated in Figure 5·3a, which presents the fraction of videos belonging to each classification over the video sequence. In Figure 5·3a the blue, green and red plots represents the fraction – marked with the y-axis values in the right – of videos from channels classified respectively as *trustable*, *neutral* and *extreme*.

The shift away from reliable videos can be seen as well in the ternary plot of Figure 5·3b. While Figure 5·3a presents the key results in a single number, the ternary plots provide a more nuanced picture. In fact, there is a simple relationship between the unreliability score and the classification proportions. In this initial ternary plot, we highlight that relation by presenting the unreliability score as a heatmap overlaid on the plot. The figure shows that at the beginning of the sequence, 59% of videos come from trustable channels, while at the end of the sequence, only 31% of videos come from trustable channels. Furthermore, the fraction of extreme videos increases over by more than a factor of six, from 1.2% to 8%.

Examining the trajectory in Figure 5·3b shows an important observation: the movement away from reliable videos is initially very fast, after which the change comes more slowly. This is seen in the variation between two consecutive observations in Figure 5·3a, but it's even more evident by the length – longer in the intial and shorter in the final observations – linking observations at Figure 5·3b. In the ternary plot at Figure 5·3b, dots, and arrows indicate the sequence progression inside a video sequence where the first and last sequence points are represented by arrows and the middle of the sequence is represented by an open circle. In this way, the length of the lines between points or arrows represents how much the fractions have changed within two observations. Therefore, we can observe that most of the changes in the proportion mix occur in the first half of the observations while changes in the second half tend to be generally smaller. This suggests that much of the significant changes in proportion mix occurs as a result of the initial recommendations, and implies that trace measurements longer than 20 steps would not likely to show vastly different results in terms of the final proportion mix.

5.4.2 Private users get less reliable recommendations

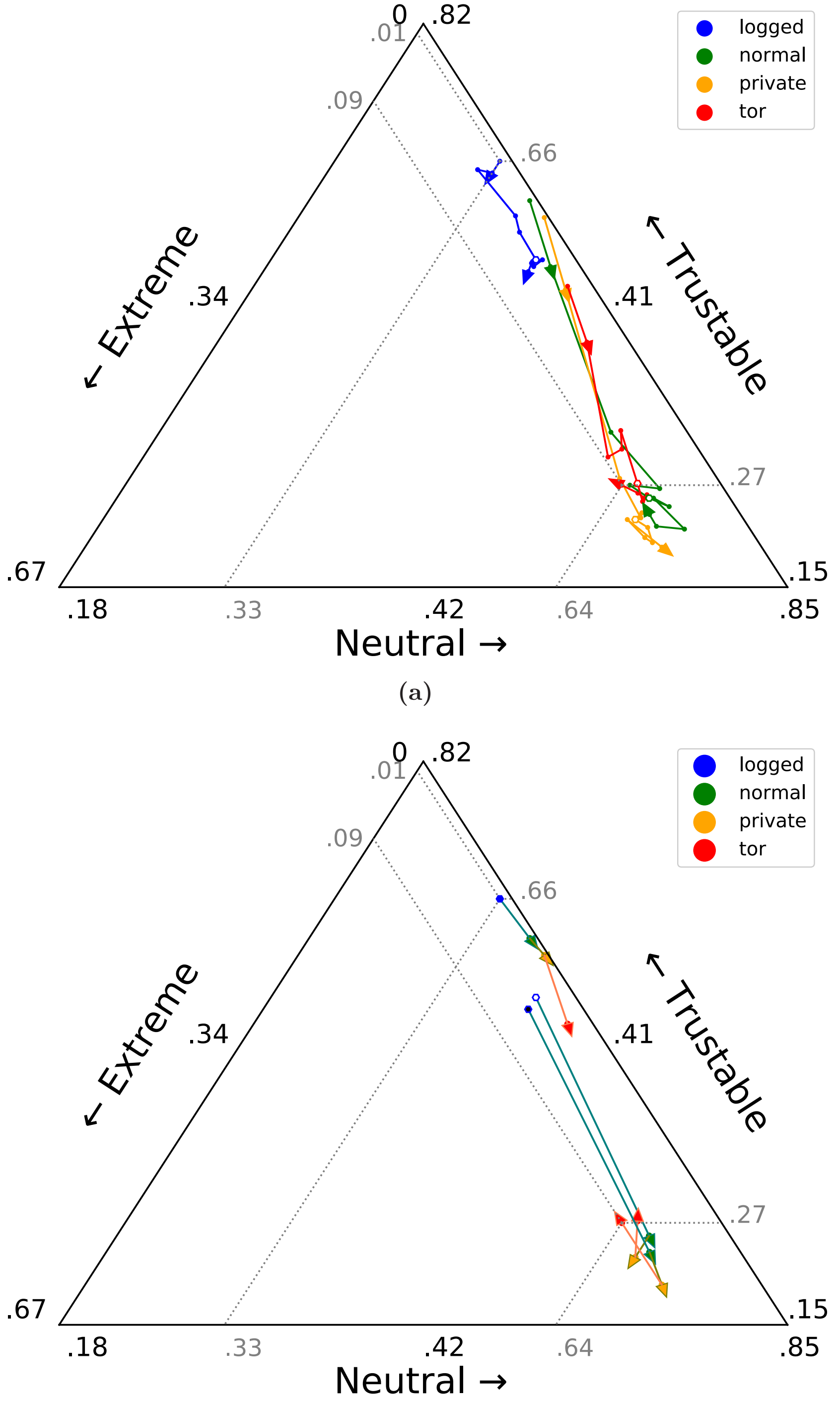
The results in the previous section are aggregated over all privacy scenarios, and hide important differences. In fact, the tendency for YouTube recommendations to lead away from reliable sources depends enormously on the user's privacy settings. We find that *privacy-seeking users are much more likely to be directed away from reliable sources and toward extreme videos.* We show this effect in Figure 5·4a where the proportions mix for each privacy scenario (*logged*, *normal*, *private* and *tor*) is shown in the ternary plot in blue, green, orange and red, respectively. We observe that all privacy scenarios show a decline in the fraction of trustable channels over the



(a)

(b)

Figure 5·3: Aggregated Shifts in Characteristics of Video Recommendations. Subfigures are: (a) classification proportions and unreliability score; (b) classification proportions.



(a)

(b)

Figure 5·4: Impact of Privacy on Shift in Recommendations. Subfigures are: (a) classification proportions shifts; (b) progression path of increases in privacy (*logged* → *normal* → *private* → *tor*) for three time samples (initial, middle and end of chain).

recommendation sequence. However the four privacy settings have considerable and important differences.

We note first of all that the overall effect of YouTube recommendations is much weaker when users are logged in – the chain path for *logged* users is much shorter than the others. Furthermore, logged in users are those with the largest proportion of trustable channels initially recommended, and the smallest decline in the fraction of trustable channels over time. For *logged* users, there is a difference of 13.1% in the trustable proportion from the initial to the end observation, while for the other privacy scenarios this difference is much larger (30.1% for *tor*, 37.3% for *normal* and, with 39.3% for *private*).

Second, we see that *tor*, *normal* and *private* settings tend to arrive at nearby endpoints, with relatively low fractions of trustable channels and high fractions of extreme channels. However, the initial recommendations provided by YouTube are quite different for *tor*, compared to *normal* and *private.* The privacy scenario *tor* starts with a low fraction of trustable channels (50.8%) while all the other settings have more than 59% of trustable channels in their initial recommendations. The lack of significant difference between *normal* and *private* suggests that if a user is not logged in, then private browsing versus normal browsing has little effect on the YouTube recommendations. This may reflect aspects of how YouTube identifies users during a browsing session.

It's important to note that the phenomenon seen in the combined data, in which the proportion mix changes fast during the first few recommendations but slower later on, is present in each individual privacy scenario as well. Figure 5·5 measures this effect by presenting the euclidian distance between ternary observation points. In this plot, we can observe that the most significant differences are between the first observations – with distances approximately ten times larger than the end observations.

This trend confirms that our conclusions regarding privacy scenarios would not likely change by observing more extended sequences of videos.

As noted, when privacy increases, there is a decrease in recommendations from reliable sources, and an increase the fraction of extreme channels. To measure this effect, in Figure 5·4b we show paths that progress along increases in privacy: from *logged*, to *normal*, to *private* to *tor*. Each path corresponds to the same time in a chain: either the initial recommendation, or the sequence midpoint (10th recommendation), or the sequence endpoint (20th recommendation). The initial points are filled in points, the sequence midpoints are open circles, and the sequence final points are black circle.

Figure 5·4b shows that early in the recommendation sequence, an increase in privacy increases the amount of extreme channels. However, for the middle and end values, there are similar amounts of extreme videos among the three privacy scenarios that doesn't disclose the user identity. Initially, the most significant increase is between *private browsing* and *tor browsing*, however by the sequence midpoints, the largest difference is between *logged* and *normal*. We also note that by the end of the recommendation sequence, the main shift in going from *logged*/*normal* to*normal*/*private* is an increase in the fraction of extreme channels.

Furthermore, the difference among *tor* and *private* reveals the role played by the IP address obfuscation. Our results show that IP address is an important factor in the initial recommendations – marked by the long arrows – but it loses its impact over time – marked by the short arrows in the middle and end observation. Additionally, the difference between *private* and *normal* suggests that when cookies are disablesd, there is a slight increase in user exposure to more extreme videos. Finally, the difference between *normal* and *logged* reveals the possible impact of knowing the user identity on YouTube's recommendations. In our results, the knowledge of

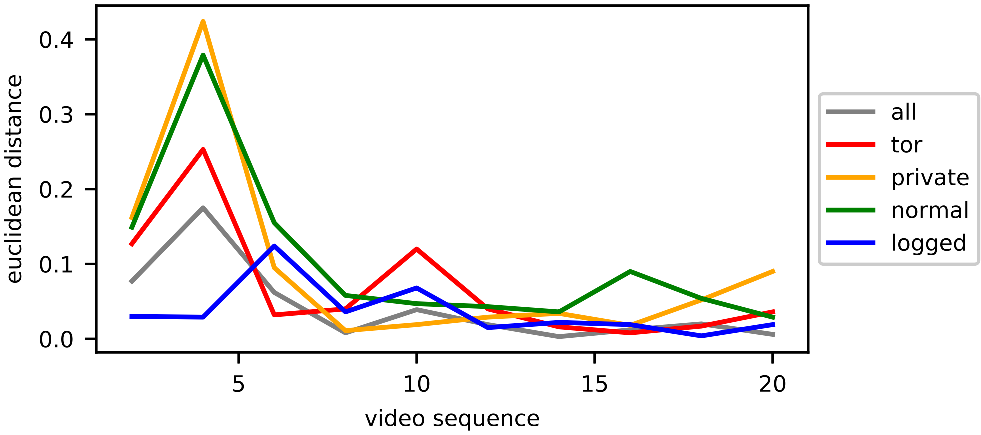


Figure 5·5: Recommendation Shift per Privacy Scenario.

user identity has a notable impact on YouTube's recommendations and significantly minimizes the exposure of the user to more extreme content.

5.4.3 YouTube more strongly recommends less reliable sources

Although the sequence of recommended channels tends away from trusted sources and tends toward extreme sources over time, we would like to assess how important the particular recommendation methods used by YouTube are to this effect. For example, it is possible that simply recommending a random set of related videos would move the user away from trusted sources.

To gauge this effect, we look at the differences between video selection of *top item* and *bottom item*. If the YouTube algorithm is actively favoring unreliable channels then we will see a greater tendency away from reliable channels when following the *top item* as compared to the *bottom item* in each recommendation list. In fact, we show that the YouTube recommender system is influencing the video outcome, that this influence is stronger in the inital part of the video sequence, and that indeed, *the recommender system is leading users to more extreme channel sources.*

In Figure 5·6 the lines and the arrows represent the difference between the initial, middle and end observation time among following the *bottom item* and the *top item*.

From the observation of *bottom item* choice, the initial, middle and end point is marked respectively by a full colored circle, a white circle with color border, a black circle with color border. First, note that the fact there is a difference between following the top or following the bottom recommendation indicates that the recommender system is actively working and not just suggesting random videos. Second, the length of the lines connecting sequences following the top and the bottom is decreasing over time. For instance, the initial arrow length – computed by the Euclidean distance between the projected points – is .049 while the end arrow length is .013. This length reduction indicates that the impact of the recommender system is considerably stronger in the initial recommendations.

Finally, the arrow direction from bottom recommendation to top recommendation indicates the prioritization of the recommender system. These directions show that the recommender system actively shifts the proportions of videos away from trustable channels in the initial and middle observation. This shift represents a reduction of 3.9% in the fraction of trustable channels for the initial observation, 2.4% for the middle observation and an increase of 0.4% of extreme channels for the final observation.

Importantly, however, this effect is not the same for all privacy settings. While the YouTube recommender system guides privacy-seeking users towards less reliable sources, when the user identity is revealed (*logged* users) the system can in fact favor more reliable sources. This effect is shown in Figure 5·7. In that plot arrows indicate the shift from following the bottom toward following the top recommendation; colors show the privacy setting adopted; and the initial, the middle and the final observations are marked as before. First, note that the initial and middle points arrows of *logged* (in blue) – with respectively length of .035 and .059 – are shorter than the other settings ones – which arrows lengths range from .002 up to .1. This shows that the

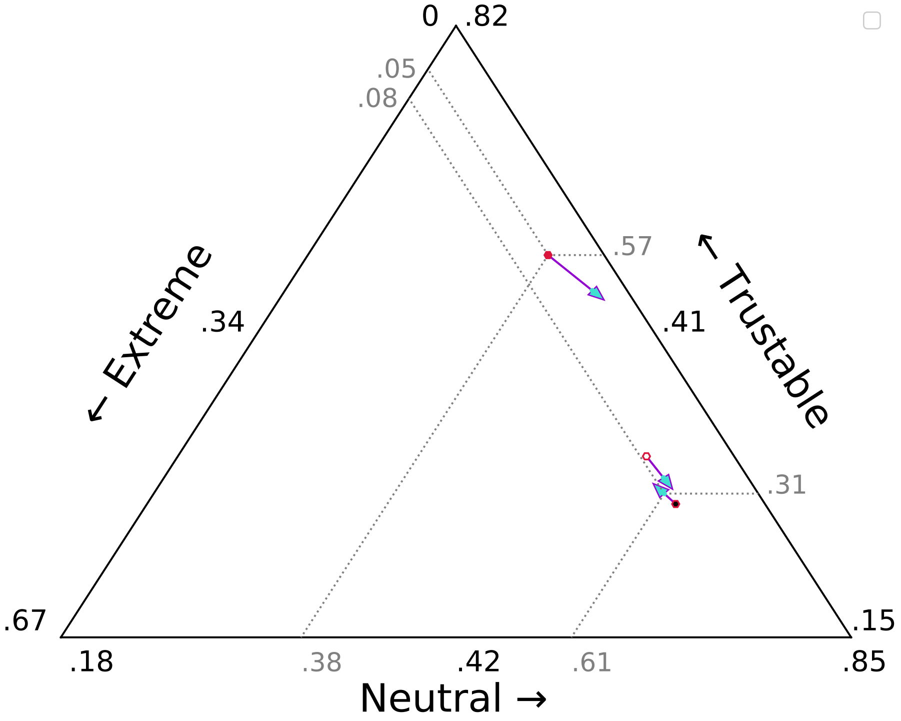


Figure 5·6: Impact of Video Selection Strategy (*bottom item* → *top item*) in time (initial, middle and end of chain)

recommender system has less effect on users whose identity was revealed. Second, observe that only for this setting, *logged*, the system effect manifest a shift towards *more trustable* and less extreme videos. For instance, in the middle observation point, there is for *logged* an increase of trustable of 7.6% and a decrease of extreme of 2.7% while the other settings for the same observation point show a a decrease of trustable of 7.4% for *normal*, 7.1% for *private* and 2.8% for *tor*. Thus, we find that the YouTube recommender system, while clearly leading privacy-seeking users away from reliable and towards extreme videos, does not have the same effect for users whose identity is known to the system.

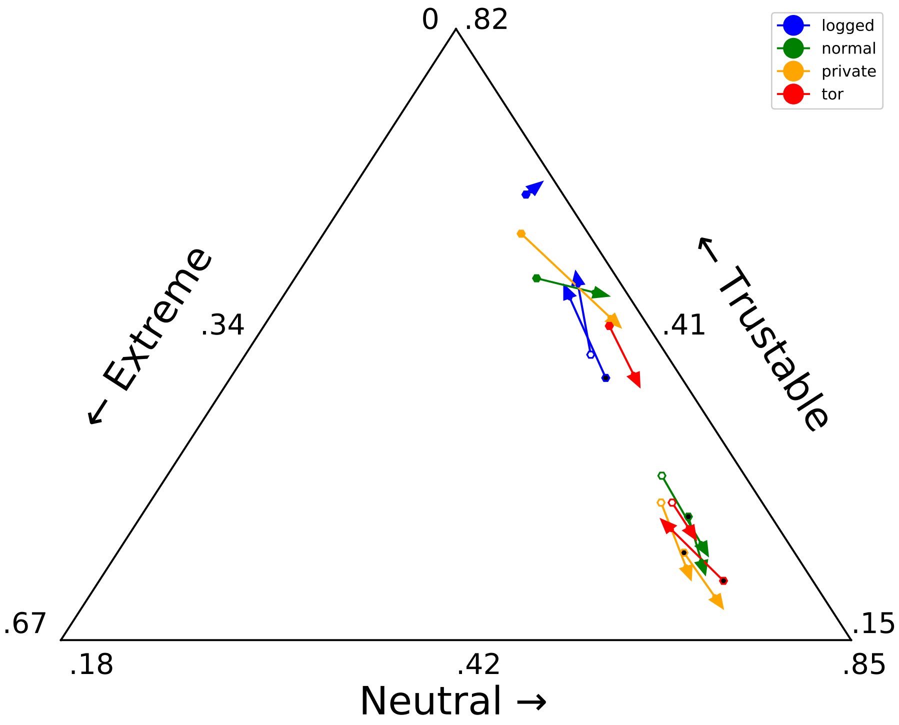


Figure 5·7: Impact of Video Selection Strategy (*bottom item* → *top item*) per Privacy Scenario.

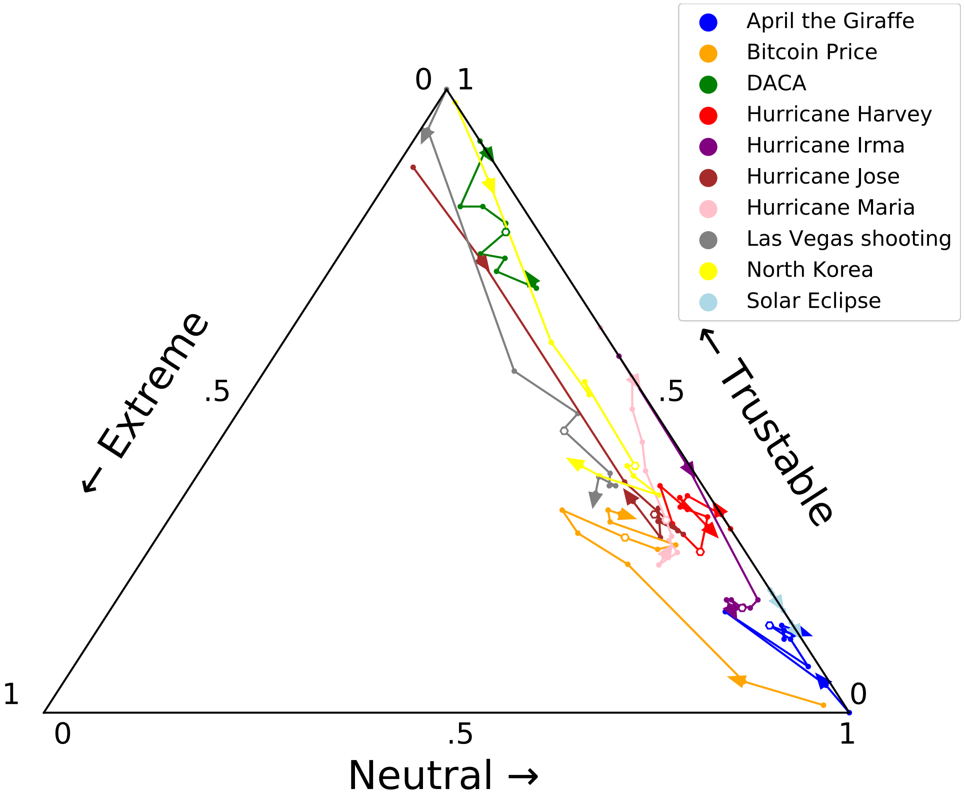


Figure 5·8: Recommendation Shift Broken Out by Search Query.

5.4.4 YouTube's recommendation effect varies depending on topic

Next, we show that YouTube's recommendation system does not affect all query topics equally. To illustrate this phenomenon, we show in Figure 5·8 the ternary plot over time for each query. The figure shows that the effect of YouTube recommendations varies considerably for different topics.

First of all, for most queries, YouTube leads users away from reliable information toward unreliable and extreme content. This is consistent with our results above. However, for some queries (*April the Giraffe* and *Bitcoin Price*) the overall movement is toward a mix of more reliable, but also more extreme content. Second, some queries are very strongly affected by YouTube recommendations (*Las Vegas Shooting* and *North Korea*) while other queries are not strongly affected (*Solar Eclipse*). In fact, the path length for the former is almost 20 times longer than that of the latter.

Inspecting the word clouds for queries we can shed light on the reasons for the differences we observe. Figure 5·9 shows the overall word clouds color coded, with word color proportional to the classification of the channel that the word comes from.

Figure 5·9 shows why some queries are relatively unaffected by the YouTube recommendation system. The small changes for queries such as *DACA* or *April the Giraffe* can be understood because the former is hard news and covered mostly by traditional news channels such as *CNN*, *ABC* or *Fox* while the latter is soft news covered mostly by animal-related channels (which are classified as neutral). Additionally, the small effect for the query *Solar Eclipse* is explained by an ambiguous results split between hard news (news about the 2017 Solar Eclipse) and soft news (the title of a song, "Solar Eclipse" by the popular singer "YoungBoy Never Broke Again").

On the other hand, some topics are much more strongly affected by the YouTube recommendation system. To unravel the reasons for different length paths, Figure 5·10 presents word clouds of specifics points in the sequence for some of the queries where each word color is associated with a RGB value proportional to it classification category fractions – red to *extreme*, green to *neutral* and blue to *trustable*. First, *Las Vegas Shooting* is the longest path of Figure 5·8 and we can observe in the time clouds of Figure 5·10 its initial dominance by traditional news channels that are replaced over time with more neutral channels and conspiracy channels (here classified as extreme). Comparing the paths of the queries *Hurricane Irma* and *Hurricane Maria* in Figure 5·8, we note that the latter is much more strongly affected than the former. Referring to Figure 5·10 , we see that *Hurricane Irma* recommendations are dominated by entertainment channels late in its sequence, shifted by a popularity bias caused by a hurricane video posted by popular YouTuber gamer "Tmartn2". On the other hand, *Hurricane Maria* presents a more balanced path between entertainment and traditional news progression. Finally, as noted above, some queries (*Bitcoin Price* and *April the Giraffe*) move *toward* more reliable sources. Figure 5·10 shows *Bitcoin Price* initially points to several independent, popular YouTubers commenting about cryptocurrencies, here classified mostly as neutral channels. However, over time, traditional news sources are recommended leading to an movement toward more reliable channels in this case.

5.4.5 YouTube's policy change did not fully remove the shift away

As we noted above, during our the data collection period YouTube implemented a change to its recommendation policy. Our analysis shows that after that change, in late January of 2019, YouTube was still leading users away from reliable sources over time. However, there was a reduction in the tendency to extreme recommendations. To illustrate that change we separate our data into two parts: data collected before February and data collected from February onward – sampling the data for the top item selection strategy, in a way that both portions of the dataset have the same number of experiments per privacy scenario and query.

Figure 5·11 plots the unreliability score for bins aggregating over 20 chains, ordered by date. We use zero on the x axis to denote first experiment of February. In this plot, we can observe that the unreliability scores for data collected before the policy change show considerable variability but do not exhibit an strong trend. However, for data collected after the policy change, there is a decreasing trend showing that less unreliable sources are recommended overall.

The change in how YouTube leads users away from realiable sources before and after the policy change is expressed in Figure 5·12a. Comparing trajectory paths we can see that in both cases, users are led away from reliable sources and toward neutral and extreme content. However, after the policy change the effect of the recommendation system is decreased overall, and fewer extreme channels are recommended. For instance the before path has an increase of 8.5% in the fraction of extreme recommendations (going from 1.2% to 9.7%) while the path after has an increase of 5.9% (going from 0.6% to 6.5%).

Comparing the effect of the recommender system before and after the policy change we find that for initial recommendations the effect was similar, favoring untrustable sources, however for later recommendations the revised recommender system shifts toward promoting more trustable channels. We see this in Figure 5·12b, which has an arrow connecting the values from the bottom item to the top item selection strategy for the initial, mid-point and end-point (marked respectively with full colored circle, a white circle with color border and a black circle with color border), comparing before and after the policy change. In this figure, while all the observed points before the change show a shift away from reliable sources, the mid and end point after the change show a tendency toward favoring more trustable sources.

Finally, we can observe in Figure 5·13 that the overall trend regarding the tradeoff of privacy and extreme content exposure is likewise present both before and after the YouTube policy change.

5.5 Summary of the chapter

In this Chapter we have presented an empirical exploration of the nature of YouTube recommendations. We developed a data collection framework that impersonates users watching a sequence of videos on YouTube, for different privacy scenarios and video selection policies. We classified the pool of recommended channels and quantified changes to nature of the recommended content over time.

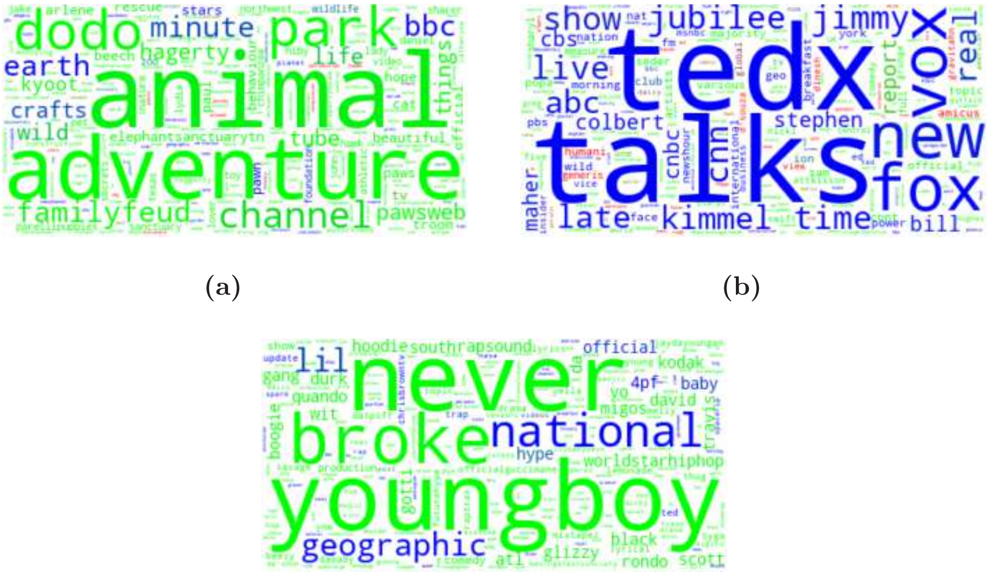
Our results show that YouTube's recommendations typically lead users away from reliable sources over time. Importantly, we pointed out where in time this shift happens, demonstrating how quickly users can be exposed to extreme information. A particular focus of our study is the tension between user privacy and extreme recommendations, and we expose the fact that privacy-seeking users are much more likely to be led away from reliable sources and towards extreme videos. Then, we show how YouTube's "lead away" effect varies according to the query topic, but that most topics we studied exhibit the effect. Finally, we find that the last changes in the YouTube recommendation policy have reduced but not yet solved the "lead away" effect.

Our work has a number of limitations that suggest further study. Any study of an existing system such as YouTube is necessarily specific to the details of that system. However, our conclusions are consistent with other studies, while also providing a quantitative view of results that are mainly qualitative in other studies. At a more detailed level, extending our approach to larger set of logged-in users would allow investigation of the extent that personalization can affect the "lead away" effect. Another desirable extension of this work would be an automatic video reliability classification system. This work contributes to this extent by providing a labeled dataset with some video's features that could be used to design and test such kind of system.

Nonetheless, our results suggest that engagement-driven recommendations, such as are used by YouTube, can have undesirable interaction with privacy-seeking users, resulting in a tendency to strongly direct such users toward unreliable information. Taken in the context of the currently-dominant business model of advertising-supported content publication, the ongoing evaluation of these effects is important for understanding their impact on society.

Moreover, observing our trade-off results between privacy-seeking and reliability, a piece of actionable advice for those users seeking for privacy but also seeking for more reliable information would be use a mock account, i.e., an account that it is not used or associated with the real user activity (e-mail, search queries, etc.) or identity – being merely used when this user is watching videos on YouTube.

Furthermore, eventually, if a real-time reliability classifier is implemented, a browser plug-in could provide an alternative option to give control to users about the displayed content. This real-time reliability classifier could be used, for instance, to flag videos according to the classification or, in case of some user adjustment or parenting control, to block or skip videos under some threshold of reliability.



(a)

(b)

(c)

Figure 5·9: .
Word Cloud of Recommended Channels' Names by Search Query. Subfigures are: (a) April the Giraffe; (b) DACA; (c) Solar Eclipse.

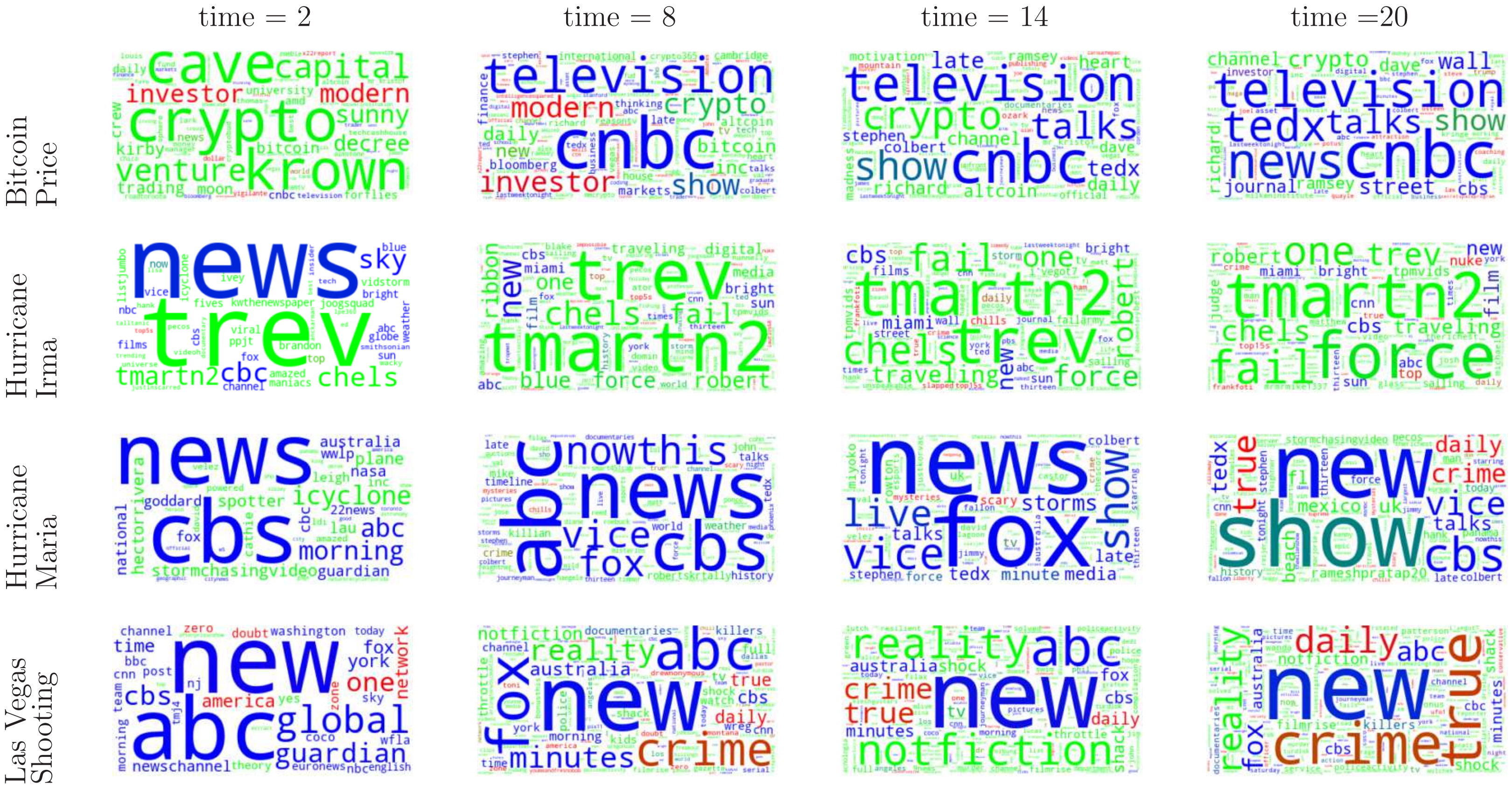


Figure 5·10: Word Cloud of Recommended Channels' Names by Search Query in Time.

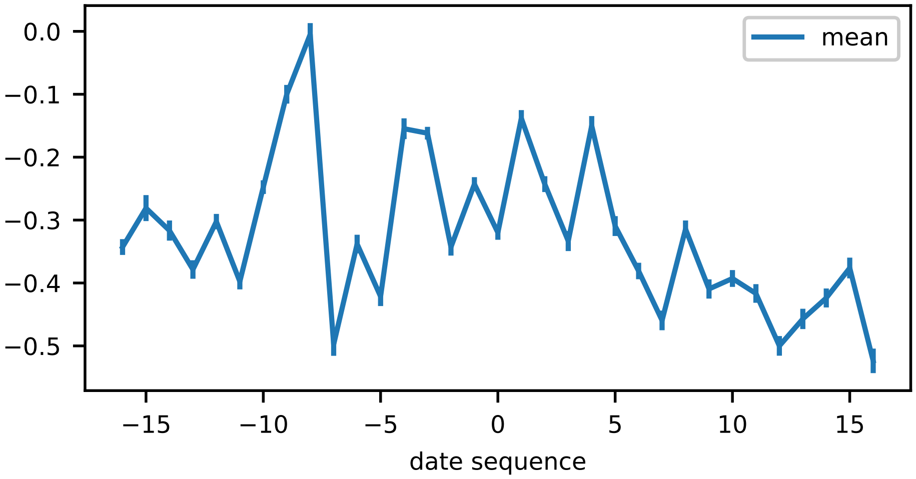
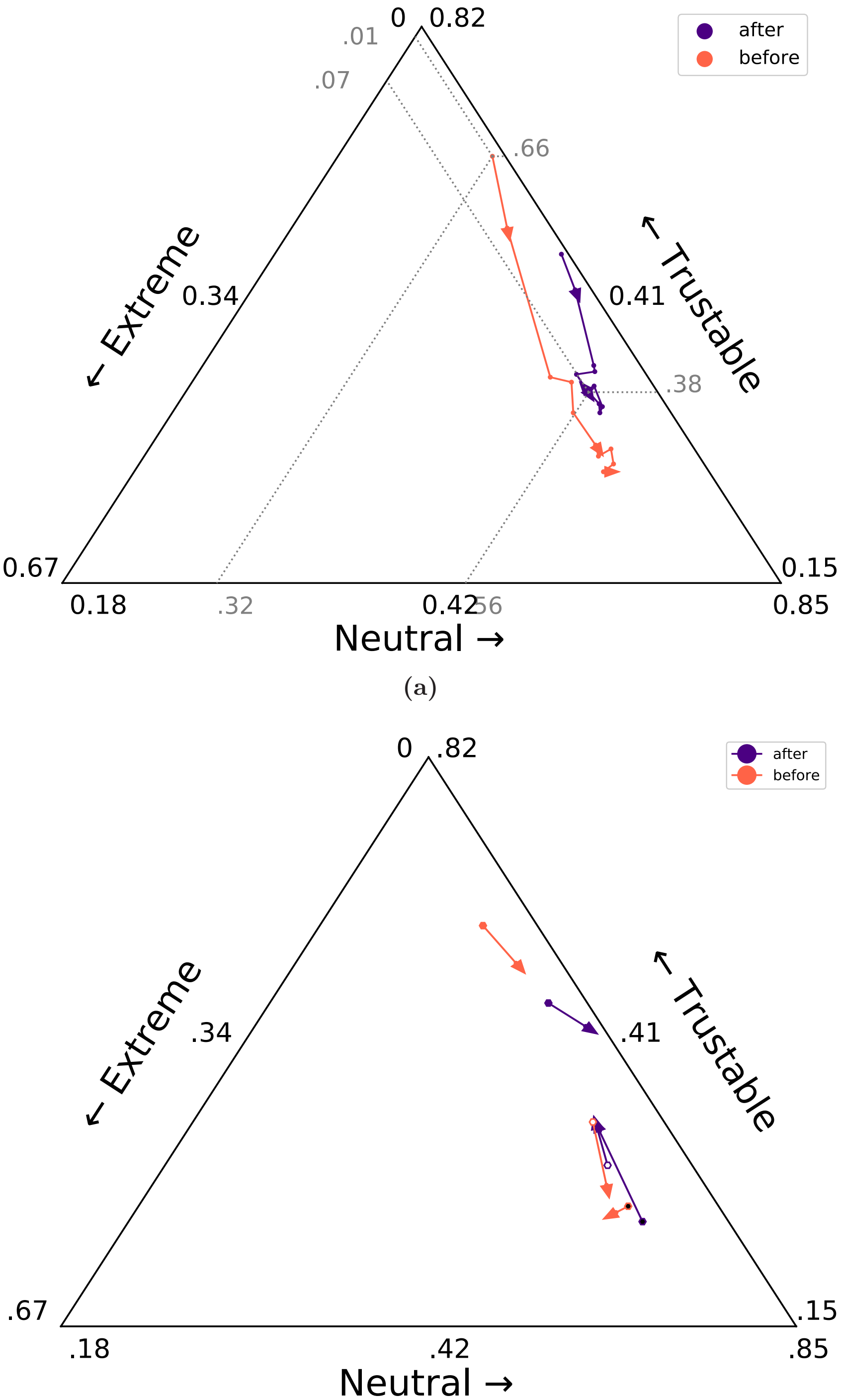


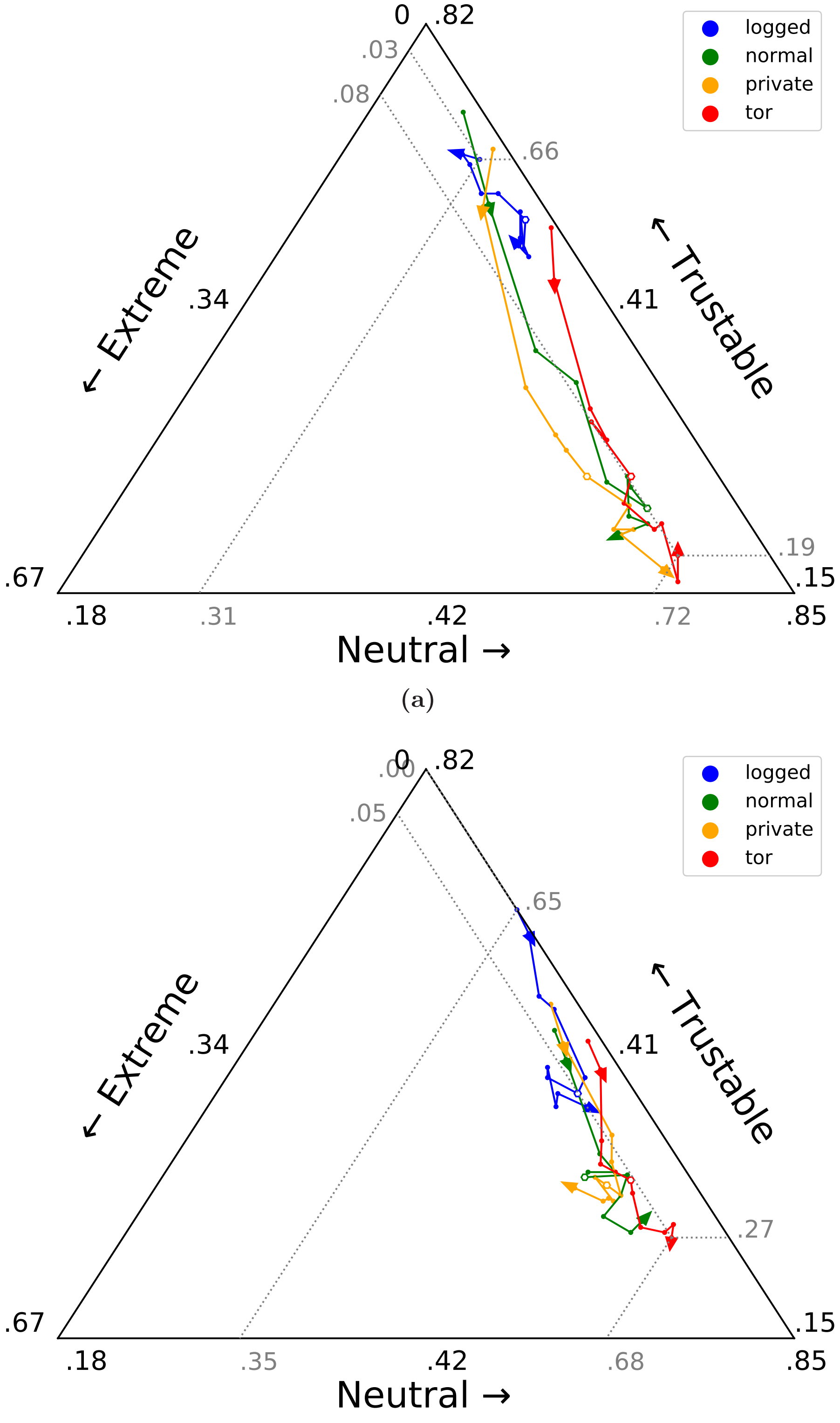
Figure 5·11: Unreliability Score Before and After YouTube Policy Change.



(a)

(b)

Figure 5·12: Recommendation Shift Before and After YouTube Policy Change. Subfigures are: (a) classification proportions shifts; (b) video selection strategy difference (*bottom item* → *top item*) in time (initial, middle and end observation).



(a)

(b)

Figure 5·13: Recommendation Shift Before and After YouTube Policy Change, by Privacy Scenario. Subfigures are: (a) before; (b) after.

Chapter 6

Conclusions

Understanding how online experiences can affect users' opinions is a multifaceted problem involving various actors and multiple dimensions of impact. In this dissertation, we contribute to the understanding of how two main drivers of impact – crowdsourced reviews and algorithmically-chosen recommendations – affect users' opinions with respect to reliability and polarization.

In the first part of this dissertation, Chapter 3, we address reliability on crowdsourced reviews by investigating how item ratings change over time and which factors play a role in those changes. We consider multiple item types across multiple ratings platforms, and use an interpretable model to break down ratings in a manner that facilitates comprehensibility. We characterize the general behavior of rating dynamics, and we quantify the contribution of factors (platform characteristics, user population, item perception, item type, and closed-loop effects) that merge to describe online rating dynamics. We show that the various kinds of dynamics observed in online ratings are primarily understood as a product of the nature of the ratings platform, the characteristics of the user population, known trends in ratings behavior, and the influence of recommendation systems. Taken together, these results provide a framework for both quantifying and interpreting the factors that drive the dynamics of online ratings allowing a more reliable assessment of a review value in time.

Next, we examine the impact of recommender systems on regarding the polarity of content made available to users and the content reliability.

In Chapter 4 we investigate dynamics of polarization by studying how the dynamical system defined by user and recommender systems behaves, as each element evolves in time. We define three metrics to understand the polarization effects – intensity, simplification, and divergence. Furthermore, we analyze the recommendation system impact over a wide variety of settings (user response, algorithms, and start point) either on linked-based systems or ratings-based systems. Our results suggest that previous studies of this problem have been too simplistic (for instance, ignoring the sensibility of systems' dynamics to certain assumptions) and that user opinions can evolve in complex ways under the influence of personalized information sources. In addition to our insights about how recommender systems can impact polarization, Chapter 4 contributes a framework – metrics and experimental setups – as a versatile tool to quantify and compare polarization effects among recommendations algorithms.

Next, in Chapter 5, we present an empirical exploration of the nature of YouTube recommendations, concentrating on socially-impactful dimensions. Our results confirm that YouTube's recommendations generally "lead away" from reliable information sources, with a tendency to direct users over time toward video channels exposing extreme and unscientific viewpoints. We show that there is a fundamental tension between user privacy and extreme recommendations. For instance, we show that in general, users who seek privacy by keeping personal information hidden, receive much more extreme and unreliable recommendations from the YouTube engine. We quantify this effect along various dimensions, including its dynamics in time, and show that the tradeoff between privacy and unreliability of recommendations is generally pervasive in the YouTube recommendation process. Further, we show that the recent efforts by YouTube to address this extremist "Rabbit Holes" problem have reduced its effect but not solved it. As a whole, Chapter 5 contributes not only with an extensive analysis of the reliability of YouTube recommendations but also with an extensible framework to analyze shifts in content reliability in sequential recommendations.

Regarding how to analyze recommender systems, our analysis on Chapter 4 and Chapter 5 have a distinct approach. At Chapter 4, we made a more open box exploration, i.e., we examine the recommender system knowing their recommendation algorithm and controlling the user response. In this way, we, therefore, had a way to investigate the system evolution in a systematic and controlled way. However, we would like to point out that our metrics can be used independently of our closed-loop system analysis – they could be used for instance analyzing input sequentially a dataset with ratings with timestamps. Furthermore, our closed-loop analysis could also be extendable into a black-box approach. In this approach, the recommender algorithm could be unknown, and the matrix completion could be used merely to estimate how the system perceives user opinions. At Chapter 5, our recommender system analysis is a complete black-box. In that analysis, we control the user privacy settings and the user watching behavior, but the process of recommendations was unknown and part of our investigation goals. Although the black-box approach is more generic – in the sense that no previous knowlegde of the system is required – is a process more challange once that, for instance, the system can be changing during the experiment collection.

Taken together, the studies presented in this dissertation shed light on important factors that affect how user opinion is shaped by online systems.

Notwithstanding our contributions, our work has limitations in the scope of our analyses. First, we would like to note that when analyzing crowdsourced reviews, our tools do not consider information from descriptive text reviews. Although how to access value and measure over time differences in text reviews add a new complexity layer to the problem and it was not addressed in this thesis, some general features extractions from text reviews could add value to our framework, for instance, by achieving a better population identification. Moreover, in reviews that support e-commerce the identification of which part of the evaluation is related to the product and which part is associated with the e-commerce experience (shopping and delivery process) would help to assess the correct value to each component.

Second, when analyzing the dynamics of polarization, although we use a simulation approach, we also encourage our metrics and tools to be used in an experimental environment. Furthermore, by showing how sensitive previous theoretical analysis is to certain assumptions, our work encourages other scholars to further the theoretical analyses including sensitive parameters such as initial state, the limit of items or avoidance of repetitive recommendations.

Finally, when studying reliability on YouTube, our analyses would be enriched by a broad set of logged users or search queries (for example, not only those related to news). Furthermore, the data from our experiments and labels associated with it can be used as a start point to the development of an automatic source classification.

Table 6.1 presents a summary of how the work present in each chapter can be extended.

The understanding of how opinion dynamics are affected is of the interest of multiple areas of studies, including social sciences. For instance, the work developed at Chapter 3, could help market professionals working with online ratings by providing a framework to understand a more holist picture of the ratings. Chapter 4 and Chapter 5, which explore the impact of recommender system on the opinion dynamics, could be utilized by professionals and researchers of social sciences that want to do, for instance, social impact assessment of information accessed through unreliable or polarized sources observing social factors such as risk assessment, cultural impact, and economic implications. Furthermore, future collaboration with such kind of researchers or professionals could help the expansion of this work for more social dimensions so far, not examed.

As a final remark, we would like to highlight that understanding how the opinion dynamics is affected in the online experience concerns our society as a whole. This understanding is complex in nature, thus and it would be enriched with a broader (more drivers of impact) and deeper (more dimensions or factors affecting the impact) exploration. Furthermore, the advent of new technologies and social demands (such as accountability, transparency, fairness) keep this challenge ongoing and in perspective of expansion.

Table 6.1: Work Extensions

Chapter 3: Unraveling the Dynamics of Online Ratings
Automatic clustering and highlight of trends; Breaking down by component with adjustment to plots to show the offset of changes components; Comparative function to contrast differences among dataset or clusters; Incorporate categorical or textual features that could be used to break down datasets into clusters;
Chapter 4: Closed-Loop Opinion Formation
The framework can extend its analyze to any recommendation algorithm – in this context, the matrix completion could be utilized merely to obtain the user opinion vector; Another definition of user opinion vector could also be applied to the metrics; Metrics can be used for analyzing sequential inputs from a dataset with timestamps (without the simulation);
Chapter 5 : How YouTube leads users away from reliable information
Our software that collects information simulation users' watching behavior can be easily extensible to another pre-conditions or further features collection; Metrics and ternary analysis – with a path connecting sequential observations – could also be applied to others sequence of recommendation with three distinct labels; The framework would gain power by implementing an automatic source classification regarding its reliability;

References

Abbassi, Z., Amer-Yahia, S., Lakshmanan, L. V., Vassilvitskii, S., and Yu, C. (2009). Getting Recommender Systems to Think Outside the Box. In *Proceedings of RecSys*, pages 285–288, New York, NY, USA. ACM.

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 17(6):734–749.

Andrew Guess, Brendan Nyhan and, J. R. (2018). Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign. https://www.dartmouth.edu/ nyhan/fake-news-2016.pdf.

Bakshy, E., Messing, S., and Adamic, L. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, pages aaa1160+.

Bell, R. M. and Koren, Y. (2007). Lessons from the Netflix Prize Challenge. *SIGKDD Explor. Newsl.*, 9(2):75–79.

Bergen, M. (2019). YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant. https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant. Accessed: 2019-04-30.

Bermingham, A., Conway, M., McInerney, L., O'Hare, N., and Smeaton, A. F. (2009). Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation. In *2009 International Conference on Advances in Social Network Analysis and Mining*, pages 231–236.

Bhawalkar, K., Gollapudi, S., and Munagala, K. (2013). Coevolutionary Opinion Formation Games. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, pages 41–50, New York, NY, USA. ACM.

Bindel, D., Kleinberg, J. M., and Oren, S. (2012). How Bad is Forming Your Own Opinion? *CoRR*, abs/1203.2973.

Chaslot, G. (2018). How Algorithms Can Learn to Discredit the Media. https://medium.com/@guillaumechaslot/how-algorithms-can-learn-to-discredit-the-media-d1360157c4fa.

Cooper, C., Lee, S. H., Radzik, T., and Siantos, Y. (2014). Random Walks in Recommender Systems: Exact Computation and Simulations. In *Proceedings of WWW*, pages 811–816, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Covington, P., Adams, J., and Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 191–198, New York, NY, USA. ACM.

Dandekar, P., Goel, A., and Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796.

Diresta, R. (2019). How Amazon's Algorightms Curated a Dystopian Bookstore. www.wired.com/story/amazon-and-the-spread-of-health-misinformation. Accessed: 2019-05-30.

Dooms, S., De Pessemier, T., and Martens, L. (2013). MovieTweetings: a Movie Rating Dataset Collected From Twitter. In *CrowdRec at RecSys 2013*.

Duan, W., Gu, B., and Whinston, A. B. (2008). The dynamics of online word-of-mouth and product sales – An empirical investigation of the movie industry. *Journal of Retailing*, 84(2):233 – 242.

Engler, T. H., Winter, P., and Schulz, M. (2015). Understanding online product ratings: A customer satisfaction model. *Journal of Retailing and Consumer Services*, 27:113 – 120.

Fandango (2016). Rotten Tomatoes. https://www.rottentomatoes.com. Accessed: 2016-09-22.

Godes, D. and Silva, J. C. (2012). Sequential and Temporal Dynamics of Online Opinion. *Marketing Science*, 31(3):448–473.

Google (2018). Year in Search 2017: See what was trending in 2017 - United States. https://trends.google.com/trends/yis/2017/US/. Accessed: 2018-04-30.

Google (2019a). Verification badges on channels. https://support.google.com/youtube/answer/3046484?hl=en. Accessed: 2019-04-20.

Google (2019b). YouTube Offical Blog: Continuing our work to improve recommendations on YouTube. https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html. Accessed: 2019-04-30.

Graells-Garrido, E., Lalmas, M., and Quercia, D. (2013). Data Portraits: Connecting People of Opposing Views. *CoRR*, abs/1311.4658.

GroupLens, R. P. (2015). MovieLens dataset ml-1m. http://grouplens.org/datasets/movielens/. Accessed: 2015-04-30.

Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., and Wilson, C. (2013). Measuring Personalization of Web Search. In *Proceedings of WWW*, pages 527–538, New York, NY, USA. ACM.

Hannak, A., Soeller, G., Lazer, D., Mislove, A., and Wilson, C. (2014). Measuring Price Discrimination and Steering on E-commerce Web Sites. In *Proceedings of IMC*, pages 305–318, New York, NY, USA. ACM.

He, R. and McAuley, J. (2016). Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 507–517, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Kevin Roose, K. C. (2019). YouTube to Remove Thousands of Videos Pushing Extreme Views. https://nyti.ms/2wNdeap. Accessed: 2019-06-05.

Kliman-Silver, C., Hannak, A., Lazer, D., Wilson, C., and Mislove, A. (2015). Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proceedings of ACM IMC*, pages 121–127, New York, NY, USA. ACM.

Knijnenburg, B. P., Sivakumar, S., and Wilkinson, D. (2016). Recommender systems for self-actualization. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 11–14. ACM.

Koren, Y. (2010). Collaborative filtering with temporal dynamics. *Communications of the ACM*, 53(4):89–97.

Krishnan, S., Patel, J., Franklin, M. J., and Goldberg, K. (2014). A Methodology for Learning, Analyzing, and Mitigating Social Influence Bias in Recommender Systems. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 137–144, New York, NY, USA. ACM.

Le, H., Maragh, R., Ekdale, B., High, A., Havens, T., and Shafiq, Z. (2019). Measuring Political Personalization of Google News Search. In *Proceedings of the 28Nd International Conference on World Wide Web*, WWW '19, New York, NY, USA. ACM.

Lempel, R. and Moran, S. (2001). SALSA: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems (TOIS)*, 19(2):131–160.

Lewis, P. (2018). Fiction is outperforming reality: how YouTube's algorithm distorts truth. https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth.

Lewis, P. and McCormick, E. (2018). How an ex-YouTube insider investigated its secret algorithm. https://www.theguardian.com/technology/2018/feb/02/youtube-algorithm-election-clinton-trump-guillaume-chaslot.

Li, W.-J., Dong, Q., and Fu, Y. (2017). Investigating the Temporal Effect of User Preferences with Application in Movie Recommendation. *Mobile Information Systems*, 2017:10.

Li, X. and Hitt, L. M. (2008). Self-Selection and Information Role of Online Product Reviews. *Information Systems Research*, 19(4):456–474.

Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80.

Liu, N. N., Zhao, M., Xiang, E., and Yang, Q. (2010). Online Evolutionary Collaborative Filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 95–102, New York, NY, USA. ACM.

Liu, Y., Cao, X., and Yu, Y. (2016). Are You Influenced by Others When Rating?: Improve Rating Prediction by Conformity Modeling. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 269–272, New York, NY, USA. ACM.

Liu, Y., Liu, Y., Shen, Y., and Li, K. (2017). Recommendation in a Changing World: Exploiting Temporal Dynamics in Ratings and Reviews. *ACM Trans. Web*, 12(1):3:1–3:20.

Maccatrozzo, V. (2012). Burst the Filter Bubble: Using Semantic Web to Enable Serendipity. In Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J. X., Hendler, J., Schreiber, G., Bernstein, A., and Blomqvist, E., editors, *International Semantic Web Conference (2)*, volume 7650 of *Lecture Notes in Computer Science*, pages 391–398. Springer.

Mäs M, F. A. (2013). Differentiation without Distancing. Explaining Bi-Polarization of Opinions without Negative Influence. *PLoS ONE*, 11(8).

McAuley, J., Targett, C., Shi, Q., and van den Hengel, A. (2015). Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 43–52, New York, NY, USA. ACM.

McAuley, J. J. and Leskovec, J. (2013). From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise Through Online Reviews. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 897–908, New York, NY, USA. ACM.

Nicas, J. (2018). How YouTube Drives People to the Internet's Darkest Corners. https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478.

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., and Cunningham, P. (2015). Down the White Rabbit Hole. *Soc. Sci. Comput. Rev.*, 33(4):459–478.

Oku, K. and Hattori, F. (2011). Fusion-based Recommender System for Improving Serendipity. In *1st International Workshop on Novelty and Diversity in Recommender System (DiveRS 2011)*.

Oram, A. (2019). What could come from the Huluization of news by Apple? https://www.linkedin.com/pulse/what-could-come-from-huluization-news-apple-andrew-oram. Accessed: 2019-04-30.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: bringing order to the web. Stanford InfoLab.

Paparrizos, J. and Gravano, L. (2016). k-Shape: Efficient and Accurate Clustering of Time Series. *SIGMOD Rec.*, 45(1):69–76.

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group USA.

Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B. (2010). *Recommender Systems Handbook*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition.

Roose, K. (2019). YouTube's Product Chief on Online Radicalization and Algorithmic Rabbit Holes. https://nyti.ms/2CLcSUT. Accessed: 2019-06-04.

Salganik, M. J. and Watts, D. J. (2009). Web-Based Experiments for the Study of Collective Social Dynamics in Cultural Markets. *topiCS*, 1(3):439–468.

Soroush Vosoughi, Deb Roy, S. A. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Spinelli, L. and Crovella, M. (2017). Closed-Loop Opinion Formation. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, pages 73–82, New York, NY, USA. ACM.

Sureka, A., Kumaraguru, P., Goyal, A., and Chhabra, S. (2010). Mining YouTube to Discover Extremist Videos, Users and Hidden Communities. In Cheng, P.-J., Kan, M.-Y., Lam, W., and Nakov, P., editors, *Information Retrieval Technology*, pages 13–24, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sweeney, L. (2013). Discrimination in Online Ad Delivery. *Queue*, 11(3):10:10–10:29.

Tufekci, Z. (2018). YouTube, the Great Radicalizer. https://nyti.ms/2GeUCDY. Accessed: 2018-03-10.

Wen, Z., Yin, W., and Zhang, Y. (2012). Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361.

Xiong, L., Chen, X., Huang, T.-K., Schneider, J., and Carbonell, J. G. *Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization*, pages 211–222.

Yang, Z., Zhang, Z.-K., and Zhou, T. (2012). Anchoring bias in online voting. *EPL (Europhysics Letters)*, 100(6):68002.

YouTube (2019). YouTube for Press: YouTube in numbers. https://www.youtube.com/intl/en-GB/yt/about/press/. Accessed: 2019-04-30.

Zhang, C., Wang, K., Yu, H., Sun, J., and Lim, E.-P. *Latent Factor Transition for Dynamic Collaborative Filtering*, pages 452–460.

Zhang, Y., Lappas, T., Crovella, M., and Kolaczyk, E. (2014). Online ratings: convergence towards a positive perspective? In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy.

Zhang, Y. C., Séaghdha, D. O., Quercia, D., and Jambor, T. (2012). Auralist: Introducing Serendipity into Music Recommendation. In *Proceedings of the ACM WSDM*, pages 13–22, New York, NY, USA. ACM.

Ziegler, C.-N., McNee, S., Konstan, J., and Lausen, G. (2005). Improving Recommendation Lists Through Topic Diversification. In *14th WWW*, Chiba, Japan. ACM.

Zuiderveen Borgesius, F. J., Trilling, D., Moeller, J., Bodó, B., de Vreese, C. H., and Helberger, N. (2016). Should We Worry About Filter Bubbles? *Internet Policy Review. Journal on Internet Regulation*, 5(1):102–114.

CURRICULUM VITAE

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]