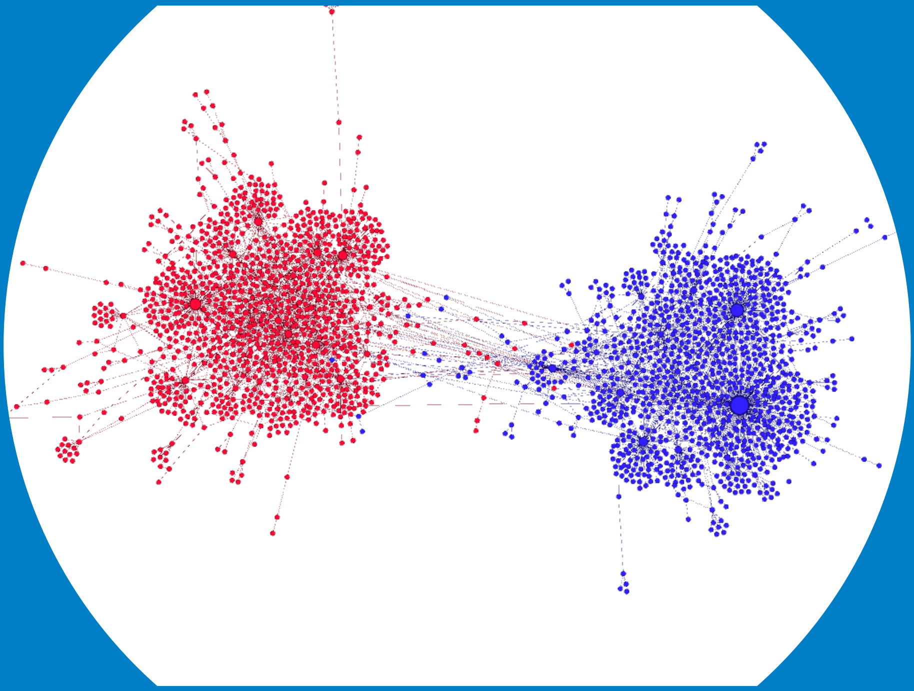
Department of Computer Sciene

Polarization on Social Media

Kiran Garimella



Aalto University

DOCTORAL
DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 20/2018

Polarization on Social Media

Kiran Garimella

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall T2 of the school on 7 February 2018 at 12:00.

Aalto University
School of Science
Department of Computer Sciene
Data Mining Group

Supervising professors
Professor Aristides Gionis,
Aalto University,
Finland

Preliminary examiners
Associate Professor Kristina Lerman,
University of Southern California,
United States of America

Professor Krishna Gummadi,
Max Planck Institute for Software Systems (MPI-SWS)
Germany

Opponents
Professor Dino Pedreschi,
University of Pisa,
Italy

Aalto University publication series
DOCTORAL DISSERTATIONS 20/2018

ISBN 978-952-60-7832-8 (printed)
ISBN 978-952-60-7833-5 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-7833-5

Unigrafia Oy
Helsinki 2018

Finland

Publication orders (printed book):
kiran.garimella@aalto.fi

NORDIC ECOLABEL
441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Kiran Garimella

Name of the doctoral dissertation
Polarization on Social Media

Publisher School of Science

Unit Department of Computer Sciene

Series Aalto University publication series DOCTORAL DISSERTATIONS 20/2018

Field of research Data Mining

Manuscript submitted 14 November 2017 **Date of the defence** 7 February 2018

Permission to publish granted (date) 22 January 2018 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

Social media and the web have provided a foundation where users can easily access diverse information from around the world. However, over the years, various factors, such as user homophily (social network structure), and algorithmic filtering (e.g., news feeds and recommendations) have narrowed the breadth of content that a user consumes. This has lead to an ever-increasing cycle where users on social media only consume content that agrees with their beliefs and hence are recommended more such content, ultimately leading to a polarized society where diverse opinions are not encouraged.

This thesis provides a broad overview of polarization on social media, along with algorithmic techniques to identify polarized topics, understanding their properties over time, and finally, to reduce polarization.

First, we provide methods to identify polarized topics automatically from social-media streams. Our methods are mainly based on interaction networks, i.e., networks of social media users, connected through certain types of interactions. We first show that polarized topics have a special bi-clustered structure in their retweet network and propose an algorithm to quantify the degree of polarization by using a random walk on this network. We then make use of sub-graph patterns (motifs) in the reply network of users to show that we can easily identify polarized topics using such patterns. Since our analysis does not use content, our methods are able to generalize to any topic, domain and language.

Next, we study the dynamic aspects of the process of polarization. We understand what happens to the interaction networks defined above in case of a sudden increase in interest of users on the topic. We then address the question on whether polarization on Twitter has increased over the last 8 years and find evidence to support that it does.

Finally, given these findings, we design algorithms to reduce polarization. We propose two approaches. In the first approach, we propose connecting users with opposing viewpoints in order to reduce polarization. Our method takes into account the users' interests and their current level of polarization to help them get connected to the people they feel comfortable in doing so. In the second approach, we take an information-diffusion route. We pose the problem of reducing polarization as a task of spreading information that reaches both sides of the polarized topic.

Keywords polarization, graph mining, social media, filter bubble, echo chambers

ISBN (printed) 978-952-60-7832-8 **ISBN (pdf)** 978-952-60-7833-5

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2018

Pages 196 **urn** http://urn.fi/URN:ISBN:978-952-60-7833-5

Preface

There are many people that I would like to thank for helping me get through with my PhD.

First and foremost, my professor Aris Gionis. I can not thank him enough for all the support through out these 4 years. Thanks for being a great mentor and friend, since the 8 years we've met. My amazing co-authors, Michael and Gianmarco – with out whom the thesis wouldn't have been possibly in the shape it is today. My long time mentor and friend, Ingmar Weber – for being a constant source of inspiration. My friends and colleagues in the department: Hristo, Geraud, Ridvan, Michael, Melik, Han, Eric, Darshan, Suhas and many others – I will definitely miss the never ending hours of lunch/dinner discussions and the short foosball stint! Finally, my family – I am forever indebted to my mother, grand parents and brother for their love and support. There were really low times during my PhD and without their unwaivering support, I could not have possibly come out successful. Thank you, అమ్మ .

Espoo, Finland, January 17, 2018,

Kiran Garimella

Contents

Preface 1

List of Publications 5

Author's Contribution 7

1. Introduction 9

1.1 Research Questions 10

1.2 Conventions 11

1.3 Contributions 12

1.3.1 Quantifying polarization 12

1.3.2 Polarization over time 13

1.3.3 Reducing polarization 13

1.4 Organization of the thesis 14

2. Background 15

2.1 What causes polarization? 15

2.1.1 Individual-level bias 15

2.1.2 Group-level bias 17

2.1.3 System-level bias 17

2.1.4 Is the internet causing polarization? 18

2.2 Quantifying polarization 19

2.2.1 Topic-level polarization 19

2.2.2 User-level polarization 21

2.3 Polarization over time 23

2.3.1 Effect of increased collective attention on polarized topics 23

2.3.2 Long-term polarization 24

2.4 Reducing polarization 25

2.4.1 Making recommendations to decrease polarization. . 26

2.4.2 Balancing information exposure 27

3. Data Collection 29

4. Quantifying polarization 33

4.1 Methods based on the Retweet network 33

4.1.1 Random-walk controversy score 34

4.1.2 User polarization 36

4.2 Methods based on the Reply network 37

4.3 Methods based on the Follow network 38

5. Polarization over time **41**
5.1 Effect of collective attention on polarized debates 41
5.1.1 Retweet network 42
5.1.2 Reply network 43
5.1.3 Content 43
5.2 Long-term Polarization 44

6. Reducing Polarization **47**
6.1 Connecting Opposing Views 47
6.1.1 Acceptance probability 48
6.1.2 User level recommendation 49
6.2 Spreading information 50

7. Limitations and future work **53**
7.1 Limitations 53
7.2 Future work 55

8. Conclusions **57**

References **59**

Publications **69**

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, Michael Mathioudakis. Exploring Controversy in Twitter. *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, 33–36, February 2016.

II Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, Michael Mathioudakis. Reducing Controversy by Connecting Opposing Views. *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, 81–90, February 2017.

III Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, Michael Mathioudakis. Exposing Twitter Users to Contrarian News. *Proceedings of the 26th International World Wide Web Conference Companion*, 201–205, April 2017.

IV Kiran Garimella, Ingmar Weber. A Long-Term Analysis of Polarization on Twitter. *Proceedings of the 11th AAAI International Conference on Web and Social Media*, 53–57, May 2017.

V Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, Michael Mathioudakis. The Effect of Collective Attention on Controversial Debates on Social Media. *Proceedings of the 10th Annual ACM Web Science Conference*, 43–52, July 2017.

VI Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, Michael Mathioudakis. Factors in Recommending Contrarian Content on Social Media. *Proceedings of the 10th Annual ACM Web Science Conference*, 263–266, July

2017.

VII Mauro Coletto, Kiran Garimella, Claudio Luchesse, Aristides Gionis. Automatic Controversy Detection in Social Media: a Content-independent Motif-based Approach. *Online Social Networks and Media Journal*, 22–31, October 2017.

VIII Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, Michael Mathioudakis. Quantifying Controversy on Social Media. *Transactions on Social Computing 2017*, Accepted for publication, July 2017.

IX Kiran Garimella, Aristides Gionis, Nikos Parotsidis, Nikolaj Tatti. Balancing Information Exposure on Social Networks. *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, 4666–4674, September 2017.

X Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, Michael Mathioudakis. Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. *Accepted for publication at the 2018 World Wide Web Conference*, Jan 2018.

Author's Contribution

Publication I: "Exploring Controversy in Twitter"

This a demo paper. The author implemented the web demo and helped in writing the paper.

Publication II: "Reducing Controversy by Connecting Opposing Views"

The author came up with the idea of the paper and designed the algorithms for reducing polarization in collaboration with the co authors. The author implemented the experiments and took a major role in writing the paper.

Publication III: "Exposing Twitter Users to Contrarian News"

This is a demo paper. The author designed and implemented the web demo. The author also helped in writing the manuscript.

Publication IV: "A Long-Term Analysis of Polarization on Twitter"

The author designed the methods in collaboration with I. Weber. The author implemented the methods and experiments proposed in this paper and played a major part in writing the manuscript.

Publication V: “The Effect of Collective Attention on Controversial Debates on Social Media”

The author worked with the co authors to come up with the methods presented in the paper. Experiments were done in collaboration with M. Mathioudakis. The author helped in writing the paper.

Publication VI: “Factors in Recommending Contrarian Content on Social Media”

The author came up with the idea of extending the above paper to a user level and designed and implemented a survey to test the hypothesis. The author collaborated with the co authors in writing the paper.

Publication VII: “Automatic Controversy Detection in Social Media: a Content-independent Motif-based Approach”

The author helped M. Coletto in coming up with the algorithms. The author performed the data collection and had a part in the writing.

Publication VIII: “Quantifying Controversy on Social Media”

The measures for identifying and quantifying controversy were a results of joint discussions with the other authors. The author collected the datasets, designed the experiments and participated in writing the manuscript.

Publication IX: “Balancing Information Exposure on Social Networks”

The author participated in discussions on developing algorithms for this paper. Experiments and datasets were conceived and performed by the author.

Publication X: “Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship”

The author came up with the idea for the paper in collaboration with the co authors and helped in writing. Experiments were designed and implemented by the author.

1. Introduction

The internet, particularly, social media, has changed the way people connect to each other and consume information over the past two decades. Social media has been playing an ever-increasing role as a facilitator for democratic discussion and debate to happen. However, social media has also been blamed for encouraging users to connect only to other like minded users and influencing what content users see, through algorithmic curation and filtering, creating "echo chambers".

With the advent of social media as a major source of news [97], it has become easier for everyone to read and share information. Even though technology has made access to diverse sources of information easy, it has not made us better at finding viewpoints that are distant from our own. Even with the availability of such information, it has become worryingly common and easy for people to restrict themselves to social circles that agree with their opinion. Search engines, social media, and news aggregators are not particularly effective at surfacing information close to our interests, but they are limited by the set of topics and people we choose to follow. Algorithmic bias and personalization accentuate these effects by providing tailored content based on a user's opinions, thus further isolating the user from a holistic view on a topic. It is easy to see that the above factors can lead to a vicious cycle, where users consume content they agree with, and social media platforms suggest content similar to that already consumed by users, thus leading users to consuming content that is restricted to a very narrow point of view. This narrow worldview breeds contempt to opposing voices, paving the path for a society that is more and more polarized.

Online polarization is very important to understand and counter as it might have adverse affects on mainstream politics, decision making in a democracy and societal life in general. Polarization can lead to users receiving biased information, which can foster intolerance to opposing viewpoints which in turn leads to ideological segregation and antagonism in mainstream political and societal issues. We've been witnessing the adverse effects of a polarized society through real world events such as the U.S. presidential elections, brexit vote, etc., where, partly due to a highly polarized environment, propaganda, and fake news have been able to make an impact.

As we observed in Publication VIII, polarized topics do not foster much discussion on Twitter. Users on opposing ends are not discussing the issue; they just ignore each other and share articles that support their view. This behavior is dangerous because discussion often helps bring out facts. Such behavior is often exacerbated by algorithmic personalization, where users are recommended content to read/share and users to follow based on their interests and previous interactions. Many users might not even be aware that they are being confined to a small set of opinions, mainly because of the opacity of the personalization algorithms. Being aware and overcoming bias in the information users consume is essential for a balanced, fair society, because social media has the power to shape voting behavior in a democratic society [97]. Furthermore, if a small set of sources (search engines or social networks) can define/decide what users see/read, it might have dire consequences in situations when a minority voice needs to be heard.

In this thesis, we present a comprehensive understanding of polarization on social media. We start by designing algorithms to automatically identify polarized topics on Twitter using patterns in various types of interactions. We design algorithms that can detect polarized topics in a domain- and language-independent manner. Next, given these polarized topics, we study their properties over time. Our work is motivated by interest in observing polarization at a societal level, monitoring its evolution to possibly understand which issues become polarized and why. We look at what happens to polarized topics in case of a sudden increase in user attention in the topic (e.g., a mass shooting incident), and study long-term trends in polarization on Twitter. Finally, we design algorithms to help alleviate this polarizaiton.

1.1 Research Questions

We base this thesis on the following research questions:

- **RQ1:** How can the emergent structure of discussions about controversial topics be measured? (Publication I, Publication VII, Publication VIII)

- **RQ2:** Can we track the evolution of these discussions and understand their dynamics over time? (Publication V)

- **RQ3:** Is the amount of polarization increasing over time, and if so, how much? (Publication IV)

- **RQ4:** Can we design algorithmic techniques to reduce polarization? (Publication II, Publication III, Publication VI, Publication IX)

Figure 1.1 shows the organization of the thesis.

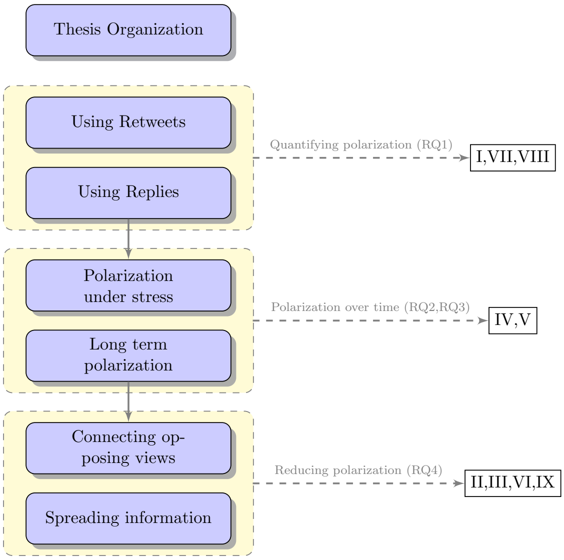

Thesis Organization
Using Retweets
Using Replies
Polarization under stress
Long term polarization
Connecting opposing views
Spreading information
Quantifying polarization (RQ1)
I,VII,VIII
Polarization over time (RQ2,RQ3)
IV,V
Reducing polarization (RQ4)
II,III,VI,IX

Figure 1.1. Thesis organization and related publications. We divide the thesis into three main components, corresponding to the four research questions.

1.2 Conventions

In this section, we define conventions that we use throughout the thesis.

Throughout the thesis, polarization refers to *political* or *social* polarization, defined as *"the act of separating people into two groups with completely opposite opinions on a topic"* (Oxford Dictionary). We are mainly interested in polarization on social media, i.e., the divergence of opinions and political attitudes to ideological extremes on social media. Consider the following question: "Should Finland welcome more refugees?" This is a contentious question to which we might get different, often conflicting viewpoints, depending on who we ask. We call this question *polarizing* and the 'topic' behind the question (refugees) as a *polarized* topic. This is because, in line with our definition, the topic separates people into two groups with opposing opinions (supporting and opposing refugees).

An important aspect in the above definition is the assumption on the existence of two opposing sides. The two sides typically correspond to either supporting or opposing a cause, or a 'yes' or 'no' (in the case of the previous example about immigration in Finland). In this thesis, we ensure that the granularity of the topics we define allows us to make a clear distinction of what the two sides represent. For instance, for the topic #obamacare, we say that users who support obamacare are one side and users who oppose obamacare are the other. On the other hand, if we consider a topic #USElections, the two groups that can be extracted from such a topic could be defined in many ways (democrats vs. republicans, clinton vs. trump supporters, etc.). We discuss the validity of the assumption about existence of two sides in Chapter 7.

In most cases discussed in the thesis, when we talk about polarization, we mean *political* polarization. The definition is general enough to accommodate other forms of polarization such as religious, cultural, and economic polarization.[1] In our experiments, we use datasets from a multitude of topics, not just confined to politics. We define polarization at a topic level and at a user level. A user is polarized if they entice opinions and information from only one side of the discussion. A topic is called polarized if there are many polarized users on each side of the discussion.

Most of the presentation in the thesis uses Twitter-specific nomenclature, e.g., retweet, reply, follow, etc. This choice is made only to simplify the explanation, since all the experiments are done using Twitter data. All our methods, however, can be generalized to other social networks like Facebook or Tumblr. Also, Twitter is a natural choice for the problem at hand, as it represents one of the main fora for public debate in online social media, and is often used to report news about current events.

We use the terms controversy and polarization interchangeably. This is because of the way we collect data to assess polarization. Our experiments mainly involve controversial discussions, which cause polarization. Hence, we are particularly interested in looking at controversial discussions as a stepping stone to understand polarization on social media, though it is not a necessity.

1.3 Contributions

The contributions we make in this thesis can be described under three main lines of work:

1.3.1 Quantifying polarization

1. Most existing work to date tries to *identify* polarized topics as case studies on a particular domain (mostly politics), either using content or social network structure. In Publication VIII, we propose an algorithm based on random walk on the retweet network, which is one of the few approaches that *quantifies* the degree of polarization of a topic. We experimentally show that our approach outperforms other competitors.

2. We then extend this method to quantify how polarized a user is. Though we are not the first to propose methods to quantify user polarization, as we see in

[1] including less-serious forms of polarization like the dress controversy `https://en.wikipedia.org/wiki/The_dress`

Chapter 4 (section 4.1.2), our method for identifying the polarity of a user is a natural extension of our method to quantify polarization of a topic.

3. In Publication VII, we build classifiers to identify polarized discussions by considering motifs in reply networks. To the best of our knowledge, we are the first to do an in-depth study of the role of reply networks in the context of identifying polarization on social media.

4. Our methods are primarily based on analyzing *interaction* networks, i.e., networks constructed using retweet and reply actions, and hence, do not depend on the content of the discussion. By virtue of this design, these methods are language- and domain-independent and hence they can be applied *in the wild* on any topic on social media (Publication I).

1.3.2 Polarization over time

1. In Publication V, we study the effects of external events on the discussion of polarized topics on social media. We collect large amounts of Twitter data pertaining to 4 long-lived polarized topics (obamacare, abortion, guncontrol and fracking) and show how different properties of these topics change with a sudden increase in attention. To the best of our knowledge, we are the first to do this for long-ranging polarized topics.

2. In Publication IV, we answer the question on whether polarization on Twitter has been increasing over the past decade. Though a lot of studies have looked at polarization in the real-world using data from surveys and voting records, there is no conclusive analysis regarding long-term trends in polarization on social media. Our study provides a long-term analysis of polarization using large-scale (over 2.5 billion tweets) and longitudinal data (around 8 years) on Twitter. We show that there is a consistent increase in polarization (around 10-20%) over the past decade on Twitter using multiple ways to measure polarization.

1.3.3 Reducing polarization

1. We design two algorithms to help reducing polarization. Although several studies have been proposed to solve the problem of decreasing polarization, there is a lack of an algorithmic approach that works in a domain- and language-independent manner, which can scale to a large number of users. Instead, the approaches are mostly based on user studies or hand-crafted datasets. To our knowledge, our work in Publication II and Publication IX is the first to offer two such algorithmic approaches.

2. Our first algorithm, in Publication II, exploits the idea of connecting users with others having an opposing viewpoint. The approach builds on existing studies from a multitude of fields including social science, psychology and human-computer interaction, to design a completely automated algorithm to reduce polarization. Most studies based on the idea of connecting opposing views focus mostly on understanding *how* to recommend content to an ideologically opposite side. Instead, the approach presented in Publication II deals with the problem of finding *who* to recommend contrarian content to.

3. Due to the scalable nature of our algorithm in Publication II and Publication VI, we were able to test it on a real-world study on Twitter consisting of almost 7 000 users. Previous studies in this area are mainly user studies involving at most a few hundred users.

4. In Publication IX, we propose an algorithm to balance information exposure and reduce polarization, in the framework of influence maximization. To the best of our knowledge, this is the first attempt to address the problem of balancing information exposure in the area of information propagation.

1.4 Organization of the thesis

This thesis follows the publication-based dissertation format of Aalto University. Due to this format, the aim of the thesis is two fold: First, to provide the necessary background in order for a reader to understand the publications and appreciate their contributions. Second, to summarize the state-of-the-art in the field, and position our contributions. We only provide high level details of the methods proposed and highlights of the results. Detailed description of the methods, proofs and evaluation can be found in the attached publications.

In particular, Chapter 2 provides a comprehensive overview of the topic of polarization from different fields including social science, political science, computer science and psychology. We first provide an overview of social theories behind polarization and then provide a detailed backgroud related to our contributions. Chapter 3 gives details on data collection and commonly used definitions. Chapter 4 summarizes the methods we propose for identifying polarized topics and quantifying their severity. Our methods encompass a wide range of user actions and interactions on social media, including retweeting, replying and following. Chapter 5 answers two questions related to the dynamics of polarization over time. In Chapter 6, we present two proposals to reduce the increasing polarization using algorithmic techniques. Finally, we conclude in Chapter 7 by presenting limitations of our methods and directions for future research.

The publications that comprise this thesis are appended in chronological order of publication.

2. Background

In the previous chapter we discussed about polarization, why it is important to study, and outlined our contributions in better understanding polarization. In this chapter, we first provide answers to the social theories that cause polarization and review existing literature to place our work in context. For each research question we pose, we review the work that has already been done in the field, and provide justification for our contributions. In particular we review work on quantifying polarization, studying the dynamics of polarization over time, and finally, reducing polarization. The study of polarization encompasses a vast amount of work from multiple fields, including social science, political science, psychology and computer science. This chapter provides a sample of studies that span these areas, and is not meant to be a thorough review.

2.1 What causes polarization?

In this section, we review some of the main factors that lead to polarization. We frame these causes in terms of well-studied social theories and define polarization as a result of various types of bias present in the society. Specifically, we define user-level biases, group-level biases and system-level biases, and show how polarization can be affected by each of those. These biases are interdependent on each other and interact in a complex way. They result in getting a user stuck in the "cycle of polarization". In particular, users make biased choices, which are reinforced when in combination with groups of like-minded users, and supported by biases from the system. Such dependence is shown in Figure 2.1.

2.1.1 Individual-level bias

First, we start with individual-level biases, which are the biases in ways users make their choices.

Cognitive dissonance. The theory of cognitive dissonance was proposed by Festinger et al. [46] and refined by Fisher et al. [48]. It refers to the phenomenon by which people experience positive feelings when presented with information that confirms that their beliefs or decisions are correct. The effects of this phenomenon extend to the level of individual media consumption behavior, for instance, the presence of opinion-reinforcing information is expected to increase the likelihood of exposure, thus reducing the exposure to a diverse source of information [57].

Homophily. Homophily is defined as the tendency of individuals to associate and bond with others who are similar to themself [74, 94]. Homophily has been measured in various facets of human behavior, including gender, race, age, status, religion, geography, etc.

On social networks, homophily leads to users connecting with (following, friending, sharing, etc) others who have similar views as their own, thus perpetuating echo chambers.

Confirmation bias. Confirmation bias is defined as the tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses.

Selective exposure. A related phenomenon is defined by the theory of selective exposure [50, 51] — which proposes the concepts of *selective exposure*, *selective perception*, and *selective retention*. It is the tendency of individuals to favor information that aligns with their pre-existing views while avoiding contradictory information.

Due to selective exposure, people keep away from communication of opposite hue. Selective perception refers to cases where, even if people are confronting unsympathetic material, they do not perceive it, or make it fit for their existing opinion. Selective retention refers to the process of categorizing and interpreting information in a way that favors one category or interpretation over another. Furthermore, they just simply forget the unsympathetic material.

Selective exposure and confirmation bias leads to biased consumption and assimilation of media choices, and hence reinforces polarized attitudes [118].

Biased assimilation. Biased assimilation [91], on the other hand, is a related phenomenon, where an individual gets exposed to information from all sides, but has the tendency to interpret information in a way that supports a pre-existing opinion. Biased assimilation is related to selective perception and retention. It is also known in part with other names such as "motivated skepticism" or "backfire effect" [114].

This phenomenon has an impact in designing systems to reduce polarization. For instance, studies have shown that the result of exposing contending factions in a social dispute to an identical body of relevant empirical evidence may be not a narrowing of disagreement but rather an increase in polarization [114].

Echo chambers. Echo chambers refer to situations where people "hear their own voice" — or, in the context of social media, situations where users consume content that expresses the same point of view that users themselves hold or express. Echo chambers have been shown to exist in various forms of online media such as blogs [59, 126], forums [43], and social-media sites [15, 66].

Echo chambers have been used to describe how information has become a partisan choice [57], and how those choices bias towards sources that reinforce beliefs rather than challenge them, regardless of the source's legitimacy [2]. However, there is contention about whether social media promotes the creation of echo chambers [15, 32].

Information overload. Information overload refers to the difficulty faced by users in understanding an issue and effectively making decisions when she has too much information about that issue [119]. The advent of internet and social media have accentuated this overload and hence this acts as a catalyst to other biases described above.

2.1.2 Group-level bias

The previous section dealt with biases at an individual level. In this section, we present group biases, stemming from collections of individuals who are similar to each other.

Social identity complexity. Social identity theory states that individuals associate themselves with social identities (race, religion, gender, class) and prefer to be part of groups that conform to those identities [115]. The social identity complexity phenomenon is similar to homophily, but at a group level.

In-group favoritism. In-group favoritism refers to favoring members of one's in-group over out-group members [39]. In the context of polarization and social media, the phenomenon is manifested by supporting and evaluating users from their own political ideology in a positive manner, while rejecting proposals by people from other ideologies.

Group polarization. Group polarization refers to the tendency for a group to make decisions that are more extreme than the initial inclination of its members [120]. These more extreme decisions are towards greater partisanship if individuals' initial tendencies are to be partisan.

2.1.3 System-level bias

Systemic biases are those that take into account biases that are not in the control of a user/group. These are biases that are perpetuated by existing institutions; they can act as a catalyst encouraging individual- and group-level biases. In the context of polarization, system-level bias could refer to two concepts:

Media bias. Media bias or operator bias refers to the perceived bias of journalists and news producers within the mass media to be biased explicitly towards a certain ideology/point of view [69]. Though media bias could be defined in a broader sense, in the context of polarization, we talk about media bias to be deliberate and explicitly favoring one side over the other. A commonly used example of media bias is the case of Fox news, which purports the conservative point of view. Studies have shown that bias in media can lead to real world

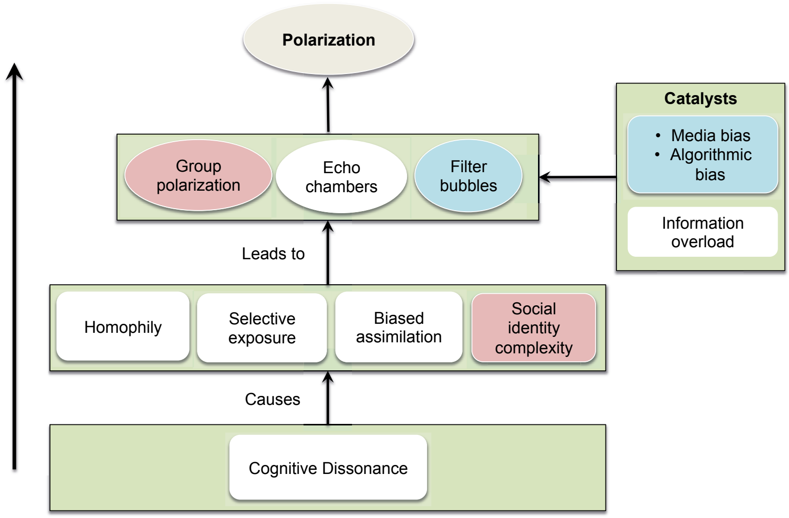


Figure 2.1. Summary of social theories and their dependencies. Individual-level (white), group-level (red) and system-level bias (blue) are colored differently.

changes in voting behavior, e.g., Dellavigna et al. [37] show that Fox News, being partisan and biased, could affect senate vote share and voter turnout. They estimate that Fox News convinced 3 to 8 percent of its viewers to vote Republican.

Algorithmic bias. Algorithmic bias refers to bias perpetuated by algorithms behind online platforms such as search engines, recommendation systems, and social networks. These biases are often invisible to users, but shape their choices. Biased algorithmic results lead to *Filter bubbles* [108], where users see information that is filtered according to their preferences, and hence reinforces their point of view.

Figure 2.1 shows the dependencies between individual-level, group-level and system-level bias and how they accentuate polarization.

2.1.4 Is the internet causing polarization?

We end this section with some discussion about whether the advent of the internet and social media platforms has actually increased polarization. Existing literature on this question has conflicting answers.

Many studies argue that the internet and social media help cause polarization because: (i) increase in available information and ultra personalized media sources — leading to people choosing agreeing information (homophily, information overload, selective exposure); (ii) increase in filtering power — people avoid reading conflicting information (confirmation bias, algorithmic filtering); (iii) increase in social feedback — homogeneity and group think reinforced (group polarization) [107, 124].

On the other hand, many studies have argued the opposite, stating that since the internet allows a wide range of choices, it helps exposing users to a much broader viewpoint [58], and facilitates cross-ideology interactions [13, 70].

A meta-analysis on whether social media encourages political participation and polarization finds evidence of a positive association between social media use and increased political participation, but questions the causal interpretation of much of the underlying evidence [22].

2.2 Quantifying polarization

We provided a basic working definition of polarization in Chapter 1, which is based on the idea of having two conflicting groups with different opinions on a topic. Polarization has been defined in many ways in different fields. Bramson et al. [24] distinguishes nine senses of polarization and provide formal measures for each one. Their main contribution is to describe polarization as distributions of attitudes/opinions. Most measures are based on ideas of quantifying the distribution of opinions, and contain methods such as spread, dispersion, fragmentation, etc.

Esteban et al. [44] propose axioms for how a measure of polarization should look like — from an economics point of view. They also propose a measure of polarization, which is an extension of the GINI coefficient [60] but also takes into account the antagonism between two sides.

For the rest of the section, we stick to our definition of polarization from Chapter 1 based on two groups of people having different opinions. We first explore topic-level polarization on social media, defined using various types of data, such as interaction networks, content and a mix of the two. Then, we look at methods that capture user-level polarization.

2.2.1 Topic-level polarization

Analysis of polarization in online news and social media has attracted considerable attention, and a number of papers have provided very interesting case studies. In one of the first papers, Adamic et al. [2] study the link patterns and discussion topics of political bloggers, focusing on blog posts on the 2004 U.S. presidential election. They measure the degree of interaction between liberal and conservative blogs, and provide evidence that conservative blogs are linking to each other more frequently and in a denser pattern. These findings are confirmed by the more recent study of Conover et al. [33], who also study polarization in political communication regarding congressional midterm elections. Using data from Twitter, they identify a highly segregated partisan structure (present in the retweet graph, but not in the mention graph), with limited connectivity between left- and right-leaning users.

The papers mentioned so far study polarization in the political domain, and provide case studies centered around long-lasting major events, such as presidential elections. In this thesis, we aim to identify and quantify polarization for any topic discussed in social media, including short-lived and ad-hoc ones (e.g., events such as #beefban[1]). The problem we study has been considered by previous work, but the methods proposed so far are, to a large degree, domain and language specific.

The work of Conover et al. discussed above [33] , employs the concept of modularity and graph partitioning in order to verify (but not quantify) controversy structure of graphs extracted from discussion of political issues on Twitter. In a similar setting, Guerra et al. [67] propose an alternative graph-structure measure. Their measure relies on the analysis of the boundary between two (potentially) polarized communities, and performs better than modularity. In a recent study, Morales et al. [98] quantify polarity via the propagation of opinions of influential users on Twitter. They validate their measure with a case study from Venezuelan politics.

Differently from these studies, our contribution consists in providing an extensive study of a number of measures, primarily based on the structure of *interactions*, and demonstrating a clear improvement over those. We also aim at quantifying polarization in diverse and in-the-wild settings, rather than carefully-curated domain-specific datasets.

In particular, we assume that polarized topics induce *retweet* graphs with clustered structure, representing different opinions and points of view. This assumption relies on the concept of "echo chambers," which states that opinions or beliefs stay inside communities created by like-minded people, who reinforce and endorse the opinions of each other. This phenomenon has been explored in many recent studies [9, 11, 49, 65, 71]. Note that the clustered structure of a retweet graph is just one condition to indicate that the topic is polarized. We can not conclude that a topic is polarized just by looking at the structure of the retweet graph. E.g. a retweet graph for promotion campaigns by different organizations might also have a clustered structure. Additional factors such as content and other interactions (reply/follow) should also be analysed to decide whether the topic is polarized or not.

Considering a different type of interaction, *conversation graphs* (reply graphs) are used to represent the dynamic nature of information and discussion threads in a network. Various studies have proposed methods to analyze reply graphs on Twitter [29, 105]. Those studies analyze various types of reply graphs, such as long path-like reply trees, large star-like trees, and long irregular trees. They also show that paths make up 60% of the reply graphs. In our work, we observe that reply graphs of Twitter discussions are composed by a majority of star-like trees. For polarized discussions, we additionally detect long trees with multiple

[1] http://www.bbc.com/news/blogs-trending-31709983

branches indicating the different threads of the discussions, e.g., see Figure 4.3 for an example visualization of a polarized discussion.

Analysis of reply graphs in rumor and misinformation spreading has shown that information flow in the network gives rise to certain types of local patterns [26, 36]. Smith et al. [117] study the role of social media in the discussion of polarized topics. They try to understand reply and retweet interactions at a user-level and conclude that users are quicker to spread information that agrees with their position more often.

The problem of detecting disagreement in reply networks was recently studied by Allen et al. [6], who use rhetorical structure features to identify disagreement. They claim that this is a difficult task, even for humans. Chen et al. [27], study when, why, and how a conversation is initiated by a controversy. Their main hypothesis is that a controversy generally brings up interest and discomfort in users, and when the former is higher, a controversy causes a conversation, while otherwise, the likelihood of starting a conversation is smaller. Supporting evidence for this hypothesis is obtained by analyzing an online news website.

A different direction for quantifying polarization was adopted by Choi et al. [28] and Mejova et al. [96]. Their method relies on text and sentiment analysis. Both studies focus on language found on news articles. In our case, since we are mainly working with Twitter, where text is short and noisy, and since we are aiming at quantifying polarization in a domain-agnostic manner, text analysis has its limitations. Nevertheless, we experiment with incorporating content features in our approach, though that is not our main focus. For details, please refer to Publication VIII.

A summary of related work along different dimensions is summarized in Table 2.1. Our contribution is shown in the last two rows of the table. We make the following distinction in existing related work:

1. Most existing work to date tries to *identify* polarized topics as case studies on a particular topic, either using content or social network structure. Our work is one of the few that *quantifies* the degree of polarization using language and domain independent methods. We show experimentally that our methods outperform others that try to quantify polarization.

2. To our knowledge, Publication VII is the first work to do an in-depth study of the role of reply networks in the context of identifying polarization in social media.

2.2.2 User-level polarization

Traditionally, the most common sources for estimating how polarized a user is comprised of behavioral data generated from roll call votes [111], co-sponsorship records [5], or political contributions [20]. These datasets were often only

Table 2.1. Summary of related work for identfying/quantifying polarization

Paper	Identifying	Quantifying	Content	Network
[28]	✓		✓	
[112]	✓		✓	
[96]	✓		✓	
[77]	✓		✓	
[122]	✓		✓	
[40]	✓		✓	
[73]	✓		✓	
[33]	✓			✓
[31]	✓			✓
[8]	✓			✓
[67]		✓		✓
[98]		✓		✓
Publication VIII		✓		✓
Publication VII	✓			✓

available for the political elite, like members of congress, and hence getting such estimates for a large population of ordinary citizens was difficult, if not impossible.

With the proliferation of social media platforms, behavioral data started being available at an individual level and researchers have tried to use such data for identifying political ideology for social media users at scale. Initial work started with supervised methods [34, 109] for predicting a (binary) political alignment of users on Twitter. Though these works report accuracies over 90%, Cohen and Ruths warn about the limitations of such approaches and their dependence on politically active users [30].

Unsupervised approaches have also been proposed, mainly based on the structure of user interests [78], social connections [14], and interactions [19, 52, 128]. The main idea behind these methods is that users typically either surround themselves (follow/friend) with other users who are similar in their ideology (homophily), or interact with others (retweet/like) similar to them.

Perhaps the closest approach to our work is by Lu et al. [92], who seek to identify the *bias* of a user on a topic by combining their retweet and content networks, where a content network is obtained based on the similarity of users tweets. The paper, however, assumes the presence of a set of labeled *bias anchors* (seed hashtags), making it not completely unsupervised. Second, fusing the content and retweet networks is somewhat arbitrary, since there is no common underlying principle that holds the two networks together and hence a graph that results from such a merger contains different types of edges (multigraph) simply merged together.

Though we are not the first to propose methods to identify how polarized a user is on social media, as we see in Chapter 4 (Section 4.1.2), our method for identifying polarity of a user is a natural extension of our method to quantify polarization.

2.3 Polarization over time

In this section, we give an overview of work done in the area of understanding the dyamics of polarization over time. We divide this into two parts: (i) understanding the properties of polarized networks in case of an external event, and (ii) general trends in polarization in the society.

2.3.1 Effect of increased collective attention on polarized topics

Most of the works mentioned in the previous section (Section 2.2) focus on static interaction networks, which are a snapshot of the underlying dynamic networks. Instead, most real world networks are dynamic and change constantly. In Publication V (and also the conference version of the work [53]), we are interested in network dynamics and, specifically, in how these networks respond to increased collective attention in the polarized topic.

Several studies have looked at how networks evolve, and proposed models of network formation [84, 85]. Densification over time is a pattern often observed [85], i.e., social networks gain more edges as the number of nodes grows. A change in the scaling behavior of the degree distribution has also been observed [3]. Newman et al. [103] offer a comprehensive review on the dynamics of networks. Most of these studies focus on social networks, and in particular, on the friendship relationship. In our work, we are interested in studying an *interaction* network, which has markedly different characteristics.

There is a large amount of literature devoted to studying the evolution of networks. For an overview, see the book by Dorogovtsev et al. [41]. However, none of these previous studies has devoted much attention to the evolution of interaction networks for controversial topics, especially when tracking topics for a long period of time.

Difonzo et al. [38] report on a user study that shows how the network structure affects the formation of stereotypes when discussing polarized topics. They find that segregation and clustering lead to a stronger “echo chamber” effect, with higher polarization of opinions. Our study examines a similar correlation between polarization and network structure, although in a much wider context, and focusing on the influence of external events.

Perhaps the closest to our work is by Smith et al. [117], who study the role of social media in the discussion of controversial topics. They try to understand how positions on controversial issues are communicated via social media, mostly by looking at user-level features such as retweet and reply rates, url sharing behavior, etc. They find that users spread information faster if it agrees with their position, and that Twitter debates may not play a big role in deciding the outcome of a controversial issue.

A few studies have examined the effects of external events on social networks. Romero et al. [116] study the behavior of a hedge-fund company via the communication network of their instant messaging systems. They find that in response to external shocks, i.e., when stock prices change significantly, the network "turtles up," strong ties become more important, and the clustering coefficient increases. In our case, we examine both a communication network and an endorsement network, and we focus on controversial, polarizing issues. Given the different setting, many of our findings are quite different.

Other works, such as the ones by Lehmann et al. [81] and Wu et al. [129], examine how collective attention focuses on individual topics or items and evolves over time. Lehmann et al. [81] examine spikes in the frequency of hashtags and whether most frequency volume appears before or after the spike. They find that the observed patterns point to a classification of hashtags, that agrees with whether the hashtags correspond to topics that are endogenously or exogenously driven. Wu et al. [129], on the other hand, examine items posted on digg.com and how their popularity decreases over time. Morales et al. [98] study polarization over time for a single event, the death of Hugo Chavez. Our analysis has a broader spectrum, as we establish common trends across several topics, and find strong signals linking the volume of interest to the degree of polarization in the discussion.

However, there are differences with Publication V:

1. We are the first to look at the dynamics of polarized topics under the influence of a sudden increase in user interest in the topic.

2. Most existing studies of similar flavor study a local topic (e.g., California ballot), over a small period of time [117], while we study a wide range of popular topics, spanning multiple years;

2.3.2 Long-term polarization

In this section, we summarize work that studies polarization a long period of time.

Lelkes et al. [83] study the impact of the introduction of broadband internet various states in the U.S. over a period of 5 years (2005-2008) and show that access to broadband increases partisan hostility. They explain that this is in part due to the consumption of partisan media. Hetherington et al. [72] study polarization of the parties over the past few decades and conclude that the increased party polarization has increased polarization in the real world. They find evidence that political parties have indeed increased in popularity by being more and more polarized.

Abramowitz et al. [1] study polarization over the past few decades using large data from the American National Election Studies and national exit polls and conclude that polarization has increased over the past decades. They suggest that, counter to popular belief that polarization turns off voters and depresses turnout, their evidence shows that polarization energizes the electorate and stimulates political participation.

Andris et al. [10] study the partisanship of the U.S. congress over a long period of time. They find that partisanship (or non-cooperation) in the U.S. congress has been increasing dramatically for over 60 years.

Finally, recently, Boxell et al. [23] studied 8 previously proposed measures of polarization and show that polarization has increased the most among the demographic groups least likely to use the Internet and social media (with age over 65 years), suggesting that the role of these factors is limited.

There are also studies that challenge the finding that polarization in real world is actually increasing.

Fiorina et al. [47] do a survey of literature on mass polarization, making a *"critical consideration of different kinds of evidence that have been used to study polarization, concluding that much of the evidence presents problems of inference that render conclusions problematic."* These results, however, have been challenged by Abramowitz and Saunders [1]. On a similar note, Prior [113] argues that *"evidence for a causal link between more partisan messages and changing attitudes or behaviors is mixed at best."*

Lelkes [82] review the different manifestations of polarization that have appeared in the public opinion literature and show that though polarization has increased, the average American has not become more polarized or ideologically consistent. They show that this increase in polarization is mainly driven by partisans, a small group of users who are politically active and increasingly dislike the other opinion.

Though a lot of studies have looked at polarization in the real world using data from surveys and voting records, there is little contribution on long term trends in polarization on social media. Our study contributes to this, by providing a long term analysis of polarization using different methods. We show that there is a consistent increase in polarization (around 10-20%) over the past decade on Twitter.

2.4 Reducing polarization

Given the ill-fated consequences of polarization on society [108, 121], it is well-worth investigating whether online polarization and filter bubbles can be avoided.

One simple way to achieve this is to "nudge" individuals towards being exposed to opposing viewpoints or read/share diverse information, an idea that has motivated several pieces of work in the literature. These related ideas on reducing polarization have been explored in various fields, including communication/media studies, political science, social science, psychology and human computer interaction (designing interfaces). Here we provide an overview and provide our contributions.

2.4.1 Making recommendations to decrease polarization.

The web offers the opportunity to easily access any kind of information. Nevertheless, several studies have observed that, when offered choice, users prefer to be exposed to agreeable and like-minded content. For instance, Liao et al. [86] report that *"even when opposing views were presented side-to-side, people would still preferentially select information that reinforced their existing attitudes."* This selective-exposure phenomenon has led to increased fragmentation and polarization online. A wide body of recent studies have studied [2, 33, 96] and quantified [4, 52, 67, 98] this divide.

Liao et al. [87, 88] attempt to limit the echo chamber effect by making users aware of other users' stance on a given issue, the extremity of their position, and their expertise. Their results show that participants who seek to acquire more accurate information about an issue are exposed to a wider range of views, and agree more with users who express moderately-mixed positions on the issue.

Vydiswaran et al. [125] perform a user study aimed to understand ways to best present information about controversial issues to users so as to persuade them. Their main relevant findings reveal that factors such as showing the credibility of a source, or the expertise of a user, increases the chances of other users believing in the content. In a similar spirit, [99] create a browser widget that measures and displays the bias of users based on the news articles they read. Their study concludes that showing users their bias nudges them to read articles of opposing views.

Graells et al. [63] show that mere display of contrarian content has negative emotional effect. To overcome this effect, they propose a visual interface for making recommendations from a diverse pool of users, where diversity is with respect to user stances on a topic. In contrast, Munson et al. [100] show that not all users value diversity and that the way of presenting information (e.g., highlighting vs. ranking) makes a difference in the way users perceive information. In a different direction, Graells et al. [64] propose to find "intermediary topics" (i.e., topics that may be of interest to both sides) by constructing a *topic graph*. They define intermediary topics to be those topics that have high betweenness centrality and topic diversity.

Based on the papers discussed above, we make the following observations:

(a) Although several studies have been proposed to solve the problem of decreasing polarization, there is a lack of an algorithmic approach that works in a domain- and language-independent manner. Instead, the approaches listed above are mostly based on user studies or hand-crafted datasets. To our knowledge, our works, Publication II,Publication IX, are the first to offer two such algorithmic approaches.

(b) Additionally, the studies discussed above on connecting opposing views focus mostly on understanding *how* to recommend content to an ideologically opposite side. Instead, the approach presented in Publication II deals with the problem of finding *who* to recommend contrarian content to. Combining the two approaches can bring us a step closer to bursting the filter bubble.

(c) The studies discussed above suggest that (i) it is possible to nudge people by recommending content from an opposing side [99], (ii) extreme recommendations might not work [64], (iii) people “in the middle” are easier to convince [87], (iv) expert users and hubs are often less biased and can play a role in convincing others [88, 125]. In the design of our algorithm in Publication II, we explicitly take into account these considerations (i)–(iv).

2.4.2 Balancing information exposure

Another direction to reduce polarization is by convincing users to read and share information from both sides. In Publication IX, we achieve this through spreading information on the network so that users have a balanced information diet. Recently, work of a similar flavor has also been done by Matakos et al. [93]. In their work, they try to find the optimal users to convince in a social network (e.g., through education, exposure to diverse viewpoints, or incentives) to adopt a more neutral stand towards polarized issues.

We now review the area of information diffusion on social networks.

Following a large body of work, we model diffusion using the *independent-cascade model* [76]. The independent-cascade model has been used extensively in different information-diffusion studies; a survey on the area is given by Guille et al. [68]. In the basic model a single item propagates in the network. An extension is when multiple items propagate simultaneously. All works that study optimization problems in the case of multiple items, consider that items *compete* for being adopted by users. In other words, every user adopts at most one of the existing items and participates in at most one cascade.

Myers and Leskovec [102] argue that spreading processes may either cooperate or compete. Competing contagions decrease each other’s probability of diffusion, while cooperating ones help each other in being adopted. They propose a model that quantifies how different spreading cascades interact with each other.

Our work is closely related to the area of *competitive information diffusion*. Most of the work in this area considers the problem of selecting the best k seeds for one campaign, for a given objective, in the presence of competing campaigns [17, 25, 104]. Bharathi et al. [17] show that, if all campaigns but one have fixed sets of seeds, the problem for selecting the seeds for the last player is submodular, and thus, obtain an approximation algorithm for the strategy of the last player. Game theoretic aspects of competitive cascades in social networks, including the investigation of conditions for the existence of Nash equilibrium, have also been studied [7, 62, 123].

The work that is most related to ours, in the sense of considering a *centralized authority*, is the one by Borodin et al. [21]. They study the problem where multiple campaigns wish to maximize their influence by selecting a set of seeds with bounded cardinality. They propose a centralized mechanism to allocate sets of seeds (possibly overlapping) to the campaigns so as to maximize the social welfare, defined as the sum of the individual's selfish objective functions. One can choose any objective functions as long as it is submodular and non-decreasing. Under this assumption they provide strategyproof (truthful) algorithms that offer guarantees on the social welfare. Their framework applies for several competitive influence models. In our case, the number of balanced users is not submodular, and so we do not have any approximation guarantees.

To the best of our knowledge, we are the only work that propose the idea of reducing polarization using the information propagation approach.

3. Data Collection

In this chapter, we first introduce some of the preliminaries of data collection on Twitter and provide definitions of some of the terms commonly used in the rest of the thesis.

Twitter is an online news and social network where users post and interact with messages posted by others. It is one of the largest social networks with over 300 million monthly active users. Though Twitter is a social network, it is mainly used also as a source of news [80], with Twitter providing the largest source of breaking news — over 40 million election-related tweets on the night of the U.S. presidential election.[1]

By default, content posted on Twitter is public and anyone can "follow" a user to receive their content. Users retweet other users for content that they agree with, and would like to spread further on the network. Retweets are not constrained to occur only between users who are connected in Twitter's social network, but users are allowed to re-post tweets generated by any other user. Throughout the thesis, we only use "pure" retweets, which do not have any additional quotes added to them (also called "quote" retweets).[2] Users can reply and mention other others, to engage in a discussion. Since most content is open on Twitter, it is one of the most accessible social networks in terms of allowing data collection at a large scale for research. Our data was collected using two main endpoints from the Twitter API.

First, the Twitter streaming API, is a 1% random sample of all tweets generated on Twitter.[3] The internet archive (`www.archive.org`), collects and archives historical samples of data from the streaming endpoint.[4] This collection dates back to 2011 and we use this data as a way to "look back" into the past. Suppose that we need to collect data about an event in 2012 (say, the Sandy hook school shooting), we first get all tweets from that time using the Archive Twitter stream. This represents only a sample of all the tweets during that time about the event. We then collect users who were actively discussing the event during that time,

[1] `http://nyti.ms/2zKTXtp` (access Nov 10, 2017).

[2] `https://support.twitter.com/articles/20169873`

[3] `https://developer.twitter.com/en/docs`

[4] `https://archive.org/details/twitterstream`

and get all the tweets of these users. Second, we use the Twitter REST API endpoint to collect data specific to a user, such as their follow network, who they retweet, tweets they post, etc.

Next, we present some common definitions that we use throughout the rest of the thesis:

Topic. A topic is operationalized as a query, and the social-media activity related to the topic consists of those items (e.g., posts) that match the given query. For example, in the context of Twitter, the query might simply consist of a hashtag. Users employ hashtags on Twitter to indicate the topic of discussion their posts pertain to. For instance, tweets corresponding to a discussion on gun control in the United States have a hashtag '#guncontrol' associated with them. For each hashtag, we retrieve all tweets that contain it and are generated during a predefined observation window. Each hashtag along with its set of related tweets define a single topic. We also ensure that the selected hashtags (topics) are associated with a large enough volume of activity.

Retweet network. After obtaining all tweets related to a specific topic (hashtag), we construct a retweet graph for the topic.[5] Each item related to a topic is associated with one user who generated it, and we build a graph where each user who contributed to the topic is assigned to one vertex. In this graph, a directed edge between two users (vertices) u and v ($u \to v$) indicates that user u retweets user v. An edge has a semantic meaning indicating endorsement, agreement, or shared point of view between the corresponding users.

Reply network. When a user publishes some content item c_i, possibly *in response to* another content item c_j authored by another user, this generates a thread of discussion. Interactions within a single thread are modeled with a content *reply tree* $\mathcal{T} = (C, R)$, where C is the set of content items in the thread, and an edge $r = (c_i, c_j) \in R$ indicates that c_i is a reply to c_j. Note that $\mathcal{T}$ is indeed a tree as each content item, except the first one (the root), is a response to exactly one other item (its parent). Additionally, the nodes of $\mathcal{T}$ are enriched with information about publishing time and authoring user. The tree $\mathcal{T}$ can be projected onto the users to model reply interactions among users. The resulting structure is a user *reply graph* $\mathcal{R} = (U, I)$, where an edge $e = (u_i, u_j) \in I$ indicates that the user u_i has replied to some content item posted by user u_j. We refer to the user who authored the first content item as *origin*.

Follow network. Users follow other users on Twitter to get access to the content produced. We construct a *topic specific follow network* consisting of follower relationships been users who discuss a topic. An edge $u \to v$ in the follow graph indicates that a user u follows user v and both users u and v were involved in the discussion of the topic.

It is commonly understood that retweets indicate endorsement, and endorsement networks for polarized topics have been shown to have a bi-clustered structure [33, 52], i.e., they consist of two well-separated clusters that corre-

[5] We use the terms network and graph interchangeably.

spond to the opposing points of view on the topic. Conversely, replies can indicate discussion, and several studies have reported that users tend to use replies to talk across the sides of a controversy [16, 90]. Follows on the other hand have a mixed role. Though, users follow other users with an opposing viewpoint, studies have shown that follow networks are usually ideologically uniform.

These different types of networks capture different dynamics of activity on Twitter, and allow us to tease apart the processes that generate these interactions.

4. Quantifying polarization

As a first step in understanding polarization on social media, we need mechanisms to detect polarized topics from large social-media streams. In this chapter, we propose two main methods to identify polarized topics and quantify the severity of polarization of the topic. Our methods are mainly based on *interaction networks*, i.e., networks of social-media users, connected through certain types of interactions.

We first show that polarized topics have a special bi-clustered structure in their *retweet* network and propose a measure to quantify the degree of polarization by using a random walk on this network. We then extend this method to identify the degree of polarization of the users involved in the discussion of the topic.

Next, we make use of subgraph patterns (motifs) in the *reply* network of users to show that we can easily identify polarized topics using such patterns. We build a classifier using various features extracted from reply networks and show that using motif features improves the classification performance significantly.

Since our analysis does not use content, our methods are able to generalize to any topic, domain and language. This is in stark difference to existing methods in this area, which are mostly case studies done on a specific topic, domain (e.g., politics), or language (english).

4.1 Methods based on the Retweet network

In this section, we explain our pipeline to identify polarized topics using the retweet network. The pipeline consists of three steps. (i) Creating the retweet graph, (ii) partitioning the graph and (iii) defining a measure to quantify polarization using this graph.

The first step in the pipeline is to construct a retweet graph. We do this as explained in Chapter 3.

Partitioning the graph. In the next stage, the resulting retweet graph is fed into a graph-partitioning algorithm to extract two partitions (as we mention in Chapter 1, we consider only polarized topics with two sides in this thesis). Intuitively, the two partitions correspond to two disjoint sets of users who possibly belong to different sides in the discussion. In other words, the output of this stage answers the following question: *"assuming that users are split into two sides according to their point of view on the topic, which are these two sides?"* If indeed there are two sides, which do not agree with each other — a polarized topic — then the two partitions should be loosely connected to each other, given the semantic of the edges. This property is captured by a measure described in the next stage of the pipeline. In principle, any graph-partitioning algorithm can be used to partition the graph. We used Metis, a spectral hierarchical partitioning algorithm [75].

We found that we can detect polarizing topics on Twitter by examining the structure of the retweet graph. Figure 4.1 illustrates the difference between retweet graphs of polarized and non-polarized topics. Intuitively, such a bi-clustered structure indicates that for a polarized topic, users are stuck in their own echo chambers and only interact with other users who agree with them. This separation is manifested in the retweet network with dense connections between users of the same side (red/blue colored nodes in Figure 4.1). We do not observe the same in the case of non-polarized topics, where the red and blue sides intersect a lot, indicating that everyone retweets everyone else.

Now, based on this observation, given a retweet graph and the two sides (obtained by clustering the graph into two partitions), we can define measures to automatically quantify the degree to which the topic is polarized. We present one such measure, based on random walks on the retweet graph in the next section. For other measures, we refer the reader to Publication VIII.

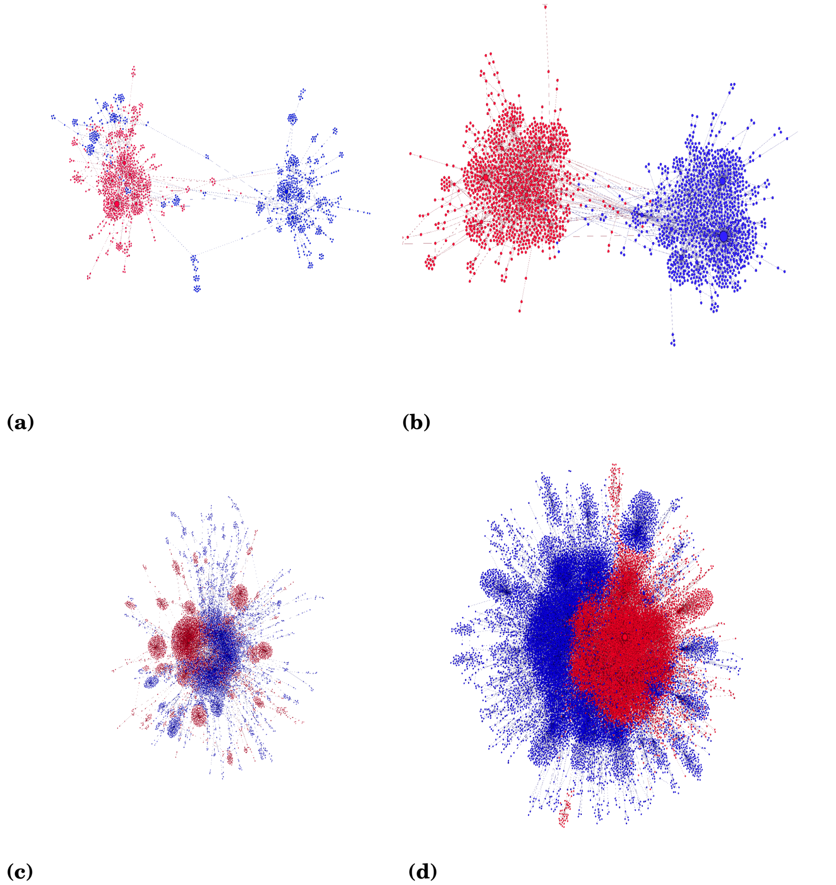
4.1.1 Random-walk controversy score

Given the retweet graph of a topic and two clusters obtained as described above, we can define a graph based measure to capture the degree of polarization of the topic using our method, called random-walk controversy score (RWC).

This measure uses the notion of random walks on the retweet graph. It is based on the rationale that, in a polarized discussion, there are authoritative users on both sides, as evidenced by a large degree in the graph. The measure captures the intuition of how likely a random user on either side is to be exposed to authoritative content from the opposing side.

We first distinguish the *k highest-degree vertices* from each partition. High degree is a proxy for authoritativeness, as it means that a user has received a large number of endorsements on the specific topic. The vertices of the retweet graph $G = (V, E)$ are partitioned into two disjoint sets X and Y, i.e., $X \cup Y = V$ and $X \cap Y = \varnothing$.

We define the *random-walk controversy* (RWC) measure as follows. *"Consider two random walks, one ending in partition X and one ending in partition Y,* RWC *is the difference of the probabilities of two events: (i) both random walks started from the partition they ended in and (ii) both random walks started in a*



(a) **(b)**

(c) **(d)**

Figure 4.1. Sample retweet graphs (visualized using the force-directed layout algorithm in Gephi). The top two are polarized topics: (a) #beefban, (b) #russia_march, while the bottom are non-controversial, (c) #sxsw, (d) #germanwings. The colors are assigned arbitrarily to the two clusters.

partition other than the one they ended in." The measure is quantified as

$$\mathrm{RWC} = P_{XX}P_{YY} - P_{YX}P_{XY}, \tag{4.1}$$

where P_{AB}, $A,B \in \{X,Y\}$ is the conditional probability

$$P_{AB} = Pr[\text{ start in partition } A \mid \text{ end in partition } B]. \tag{4.2}$$

The aforementioned probabilities have the following desirable properties: (i) they are not skewed by the size of each partition, as the random walk starts with equal probability from each partition, and (ii) they are not skewed by the total degree of vertices in each partition, as the probabilities are conditional on ending in either partition (i.e., the fraction of random walks ending in each partition is irrelevant). RWC is close to 1 when the probability of crossing sides is low, and close to 0 when the probability of crossing sides is comparable to that of staying on the same side.

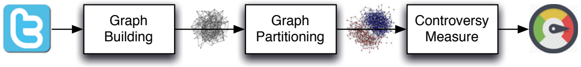


Figure 4.2. Pipeline for quantifying polarization.

This measure can be computed and updated efficiently via two personalized PageRank [106] computations, where the probability of restart is set to a random vertex on each side, and the final probability is taken by considering the stationary distribution of only the high-degree vertices. For more details, please refer to Publication VIII.

These methods, described in the sections above can be captured in a pipeline, that, given a stream of tweets, can give us as output a polarization score, obtained from RWC. This pipeline is shown in Figure 4.2. We first filter tweets by a topic and construct a graph. Then we partition the graph into two sides using an off the shelf graph-partitioning algorithm. Finally, we use RWC (or other methods) applied on this graph to obtain a score for the severity of polarization for the topic.

4.1.2 User polarization

The previous sections present measures to quantify the degree of polarization of a topic. In this section, we propose a measure to quantify the degree of polarization of a single user in the graph. We denote this score as a real number that takes values in $[-1, 1]$, with 0 representing a neutral score, and ± 1 representing the extremes for each side. Intuitively, the polarization score of a user (also called polarity score or leaning) indicates how "biased" the user is towards a particular side on a topic. For instance, for the topic 'abortion', pro-choice/pro-life activist groups tweeting consistently about abortion would get a score close to -1/+1 while average users who interact with both sides would get a score close to zero. In terms of the positions of users on the retweet graph, a neutral user would lie in the "middle", retweeting both sides, where as a user with a high polarity score lies exclusively on one side of the graph.

We can make a simple change to the above random-walk measure (RWC) to define the polarization score for each user in the graph. The score is based on the expected *hitting time*[1] of a random walk that starts from the user under consideration and ends on a high-degree vertex on either side. Typically, in a retweet graph, high-degree vertices on each side are indicators of authoritative content generators because highest degree users means that their content gets retweeted many times. We denote the set of the k highest degree vertices on each side by X^+ and Y^+. Intuitively, a vertex is assigned a score of higher absolute value (closer to $+1$ or -1), if, compared to other vertices in the graph, it takes a very different time to reach a high-degree vertex on either side (X^+

[1] Hitting time of random walk (h_{uv}) is the expected number of steps in a random walk starting at a vertex u to reach vertex v.

or Y^+) (in terms of information flow). Specifically, for each vertex $u \in V$ in the graph, we consider a random walk that starts at u, and estimate the expected number of steps, l_u^X before the random walk reaches any high-degree vertex in X^+. Considering the distribution of values of l_u^X across all vertices $u \in V$, we define $\rho^X(u)$ as the fraction of vertices $v \in V$ with $l_v^X < l_u^X$. We define $\rho^Y(u)$ similarly. Obviously, we have $\rho^X(u), \rho^Y(u) \in [0,1)$. The polarization score of a user is then defined as

$$\mathrm{RWC}_{user}(u) = \rho^X(u) - \rho^Y(u). \tag{4.3}$$

Note that the score $\mathrm{RWC}_{user}(u)$ takes values between -1 and 1. A vertex that is close to high-degree vertices X^+, compared to most other vertices, will have $\rho^X(u) \approx 1$; on the other hand, if the same vertex is far from high-degree vertices Y^+, it will have $\rho^Y(u) \approx 0$; leading to a polarization score $\mathrm{RWC}_{user}(u) \approx 1 - 0 = 1$. The opposite is true for vertices that are far from X^+ but close to Y^+; leading to a polarization score $\mathrm{RWC}_{user}(u) \approx -1$.

We also propose a variant of the user polarity score based on a modified version of personalized pagerank used for RWC. Please refer to Publication VIII for details.

4.2 Methods based on the Reply network

Retweets are a sign of endorsement; that is, they only capture the existence of a positive interaction. As Figure 4.1 shows, for polarized topics, users retweet others that they agree with and ignore users who are not from the same side. In this section, we explore whether there are other types of interactions between opposing groups in a polarized discussion.

Users on Twitter usually reply or mention other users to engage in a discussion/conversation. It has been shown [117, 16] that users on Twitter do not retweet others from opposing sides, but reply to them. To this end, we looked the structure of interactions in who replies to whom on Twitter to detect if a conversation is polarized. Note that different from the above section, here, our unit of measurement of polarization is a *conversation* and not a *topic*. A topic can have multiple conversations, but a conversation will be mostly about a single topic. Thus, we can easily extend these methods to apply to a topic level.

First, we construct the *reply tree* and the *user reply graph* as detailed in Chapter 3. Our main hypothesis in looking at replies is that the structure of the reply tree can be characterized by simple *motifs* of local user interactions that can be effectively exploited to distinguish between polarized and non-polarized content. Figure 4.3 shows the difference between reply trees for polarized and non-polarized tweets for the same origin user, @realDonaldTrump.

In addition to local motifs, we also explore whether other features, including network structure, content propagation, and temporal features can be used to distinguish polarized tweets in Publication VII. A summary of all the features

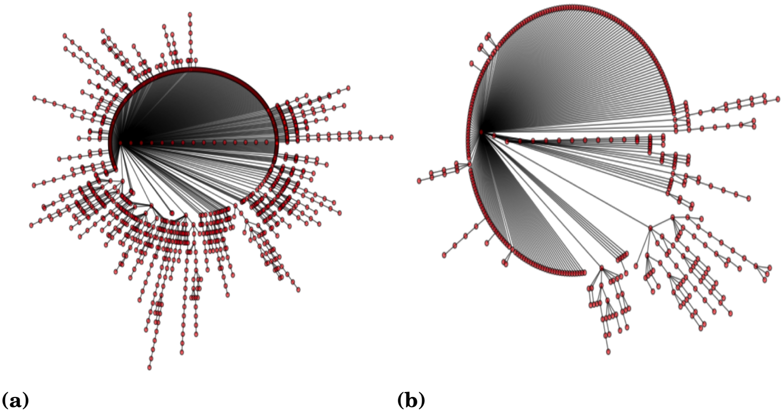


Figure 4.3. Sample reply trees for polarized and non-polarized conversations. All dots start with a root tweet and each subsequent edge in the tree is a reply.

we considered are shown in Table 4.1.

Our results show that in most cases polarized conversations arise when users participate to discussions beyond their social circles. This means that it is less likely to have polarized discussions among friends. Our classifiers using motif patterns can achieve 85% accuracy in the task of identifying polarized discussions, with an improvement of 7% compared to a baseline classifier using just structural, propagation and temporal features.

Also, as the proposed motifs can be easily extracted from any reply tree or subtree, we experimented with the use of such patterns in monitoring the evolution of discussions and sub-discussions over time. The idea behind this is that, even though a discussion might start as non-polarized, it might become polarized over time due to the way certain users reply in the discussion. Indeed, using our approach we found that a topic of discussion develops over time changing its level of polarization depending on different sub-topics or on external events (e.g., news). We found that about 7% of the direct-reply sub-trees of a non-polarized tweet become polarized.

4.3 Methods based on the Follow network

In the previous sections, we have seen methods that use interaction networks to identify polarization. In this section, we will briefly describe other methods that we proposed to make use of a different type of network — the *follow network*, also called the *social network* — to identify polarization.

We observed that using the topic-specific follow network, described in Chapter 3, works decently well in detecting polarized topics, though the signal is not as clear as it is for retweet networks. Figure 4.4 shows sample follower graphs

Table 4.1. Summary of all features used in Publication VII.

structural	num of nodes in $\mathcal{T}$
	num of edges in $\mathcal{T}$
	Avg. degree in $\mathcal{T}$
	Avg. degree in $\mathcal{R}$
propagation	Avg. cascade depth in $\mathcal{T}$
	max cascade depth in $\mathcal{T}$
	max size subtree in $\mathcal{T}$
	Max. relative degree
	Max. degree in $\mathcal{T}$ / root degree in $\mathcal{T}$
	2nd max degree in $\mathcal{R}$
temporal	Avg. inter-reply time
	max reply time
	min reply time
	% replies in 1h
dyadic motifs	7 2-node motifs
triadic motifs	20 3-node motifs
	Triangles ratio

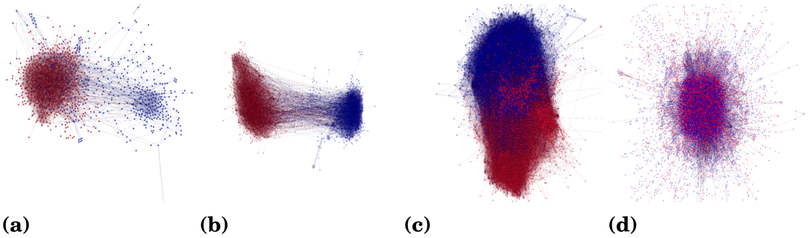


Figure 4.4. Sample follow graphs for polarized topics, (a) #beefban, (b) #russia_march, and non-polarized topics, (c) #sxsw, (d) #germanwings.

for polarized and non-polarized topics. We can clearly see that the two sides (red and blue) are separated for polarized topics, but the separation is not as clean as in Figure 4.1.

Another approach to make use of the follower network is by simply counting the number of users with a known polarity followed on a specific side of the discussion. This gives us the probability of a user to follow a certain side. Examples of users with known polarity could include, left and right leaning media outlets like @dailykos or @breitbart; or how @barackobama and @realDonaldTrump lean on the topic of immigration.

As, due to sparsity, following only a single user from one of the two sides is not necessarily a strong signal for polarization, we decided to apply a Bayesian methodology. Before observing any evidence, we gave each following user a uniform prior probability to follow a set of seed users — users with known leaning towards the polarized topic.[2] Concretely, we used a beta distribution

[2] In our case, we used a list of U.S. politicians who are either democrats or republicans.

with a uniform prior ($\alpha = \beta = 1$), where α measures the level of polarization for one side and β for the other side.

Then every follow to either side increases the count for that side by +1, basically simulating a repeated coin toss where we are studying the bias of the coin. As the beta distribution is the conjugate prior of the binomial distribution, we might obtain something like $\alpha = 4$, $\beta = 2$ for a user who (mostly) supports the first side. The mean of the beta distribution, and hence the "level of polarization" l of the following user, is defined as $l = \alpha/(\alpha + \beta)$. We defined the polarization p as $p = 2 \cdot |0.5 - l|$, giving a measure between 0.0 and 1.0 measuring the deviation from a balanced leaning. We use this measure of user polarity to estimate the increase in polarization over the last decade on Twitter in Publication IV. Details on the application of this measure are given in Section 5.2.

Using follow information does not add much value and it is practically hard to obtain due to stricter restrictions on the Twitter API for getting follower data. Retweet information, as we've explained above, is easier to obtain using the Twitter API and has a cleaner signal in identifying polarized topics.

5. Polarization over time

In the previous chapter we tried to automatically identify polarized topics by representing them as networks. Can we understand what happens to these networks over time, and how they evolve, especially when there is an external event (e.g., a mass shooting) that leads to a sudden increase in user interest in the topic? Can we use the methods proposed in Chapter 4 to answer if polarization is increasing on Twitter over the years?

This chapter provides answers to these questions. We divide the research questions into two parts. In the first part, we look at the effect of a sudden increase in collective attention on the structure of the network and the discussions of polarized topics. In the second part, we answer whether polarization on Twitter has increased over the last decade.

5.1 Effect of collective attention on polarized debates

We study the evolution of long-lived polarized debates as manifested on Twitter from 2011 to 2016. Specifically, we explore how the structure of interactions and content of discussion varies with the level of collective attention, as evidenced by the number of users discussing a topic. First, we build two types of interaction networks, a retweet network and a reply network, as described in Chapter 3.

Let us now consider the temporal dynamics of these interaction networks. Given the traditional daily news reporting cycle, we construct these networks with the same daily granularity. This high resolution allows us to easily discern the level of interest in the topic, and possibly identify spikes of interest linked to real world external events, as shown in Figure 5.1. The figure shows the daily number of active users discussing 4 long-term polarized topics: abortion, guncontrol, obamacare and fracking. Spikes in the volume of users typically correspond to external events that increase the public attention on the topic, as, for instance, discussions about ‘gun control’ often erupt after a mass shooting.

Core users. As shown in Figure 5.1, there is a base mass of users who are always active on the topic (shown with the black mass at the bottom of each sub plot). These users are typically topic specific, dedicated accounts, which

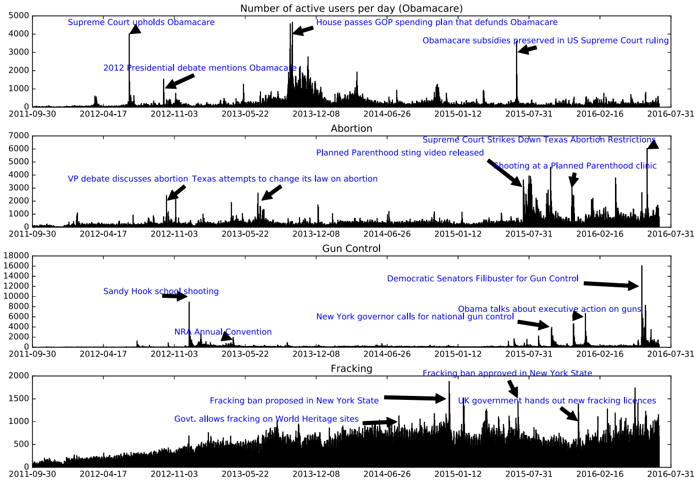


Figure 5.1. Daily trends for number of active users for the four polarized topics under study. Clear spikes occur at several points in the timeline. Manually chosen labels describing related events reported in the news on the same day are shown in blue for some of the spikes.

tweet only about that topic, for instance, gun-rights groups, gun-control advocacy groups, etc. Therefore, to understand the role of these more engaged users, we define the *core* network as the one induced by users who are active for more than 3/4 of the observation time. Nodes of a network that do not belong to the core are said to belong to the *periphery* of that network.

We now present some of the results of the effect collective attention on (i) the retweet network (Section 5.1.1), (ii) reply network (Section 5.1.2) and (iii) content (Section 5.1.3).

5.1.1 Retweet network

When an external event happens, we investigate changes to the retweet network. Usually, the core set of users are actively discussing the topic. When a sudden external event happens, we observed the following changes to the retweet network:

- New users enter the discussion — these are users who join the core users (called the periphery) and start discussing the topic. This is expected given that when an external event happens, there is media coverage on the event and normal users join the discussion.

- Most retweets are to an existing set of core users — the new users who join the discussion disproportionately retweet the existing core set of users. This can also be understood as a way that the core users becoming the “authoritative voice” during the event and other users reinforcing their voice.

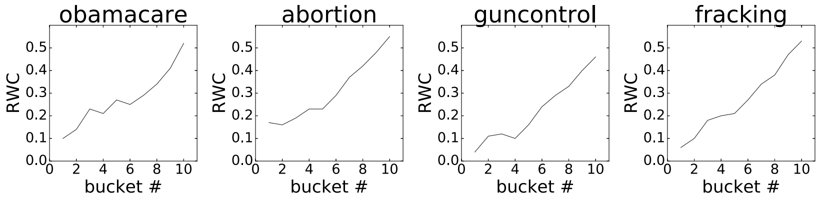


Figure 5.2. RWC score as a function of the activity in the retweet network. An increase in interest in the controversial topic (x-axis) corresponds to an increase in the controversy score of the retweet network.

- Across-side retweets decrease and within-side retweets increase — during an event, the polarization of the retweet network increases, which is manifested by a considerable increase in endorsement of users who belong to the same side and a decrease in endorsements across sides.

  Figure 5.2 shows the RWC score as a function of the quantiles of the network by volume. The x-axis shows volume of users bucketed into 10 buckets. This trend suggests that increased interest in the topic is correlated with an increase in controversy of the debate, and increased polarization of the retweet networks for the two sides.

5.1.2 Reply network

As we saw in Section 4.2, reply networks for polarized topics consist of cross edges, i.e., edges that go between the two sides. The main change that occurs in the dynamics of reply networks in case of an external event is:

- There is an increase in the amount of discussion. The discussion is mainly due to across-side edges — users reply more to other users from the other side. This, in addition to the above observation of decreasing cross side retweets indicates that the reply network might be used for disagreeing with the other side.

5.1.3 Content

Let us now switch our attention to the content being discussed and the impact of the increased collective attention on the content being generated. We measure differences in content using the differences between the unigram word distributions for the two sides before and during an event. The main observation is that the Jensen-Shannon divergence [89] between the two sides decreases. This decrease indicates that the lexicon of the two sides tends to converge. The cause of this phenomenon might be the participation of casual users to the discussions, who contribute a more general lexicon to the discussion. Alternatively, the

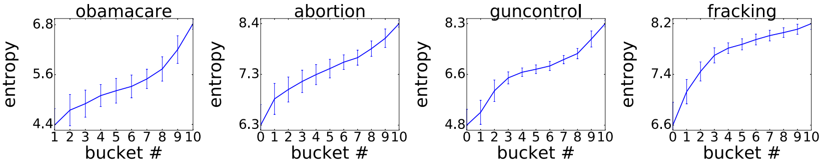
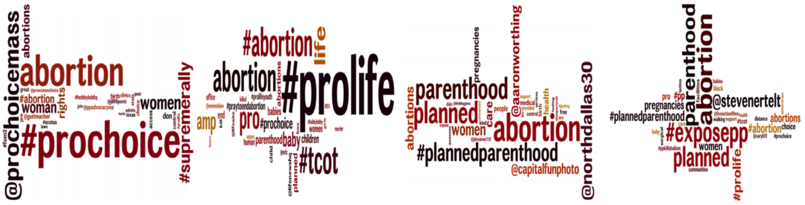


Figure 5.3. Entropy of the distribution over the lexicon for one side of the discussion as a function of the activity in the network (the other side shows similar patterns). As the interest increases, the entropy increases, thus indicating the use of a wider lexicon.



(a) **(b)** **(c)** **(d)**

Figure 5.4. Word clouds of content in a discussion on abortion before (a,b) and after (c,d) an event.

cause might be in the event that sparks the discussion, which brings the whole network to adopt similar lexicon to speak about it, i.e., there is an event-based convergence.

To further examine the cause of the convergence of lexicon, we report the entropy of the unigram distribution. Figure 5.3 shows that the entropy for one of the sides increases as interest increases (results for the other side show similar trends). Thus, we find that the lexicon is more uniform and less skewed, which supports the hypothesis that a larger group of users brings a more general lexicon to the discussion, rather than the alternative hypothesis of event-based convergence.

Figure 5.4 shows a visual example in case of the topic abortion. We can clearly see in Figure 5.4 (a,b) that there were two distinct groups — prochoice and prolife, where, as we see in Figure 5.4 (c,d), after the event the discussion is more uniform and specific to the event (planned parenthood).

A complete set of other measures we tried and results we obtained are discussed in Publication V.

5.2 Long-term Polarization

In the previous section, we looked at local changes in behavior of polarized topics when there is an external event. In this section, we want to answer the question on change of polarization at a global level, over a long period of time. Twitter and other social networks have only been in existence since the last decade or so. There have been studies using real world surveys to quantify the increase in polarization over time [1, 113], though there are other studies that conflict this

Table 5.1. U.S. seed accounts with known political leaning. Top: political candidates and parties. Bottom: partisan media outlets.

Political accounts	**Side**
barackobama,joebiden,timkaine,hillaryclinton, thedemocrats	left
realdonaldtrump,mike_pence,mittromney,gop, speakerryan,senjohnmccain,sarahpalinusa	right
Media outlets	**Side**
npr,pbs,abc,cbsnews,nbcnews,cnn,usatoday, nytimes,washingtonpost,msnbc,guardian, newyorker,politico,motherjones,slate, huffingtonpost,thinkprogress,dailykos,edshow	left
theblaze,foxnews,breitbartnews,drudge_report, seanhannity,glennbeck,rushlimbaugh	right

conclusion [47].

We test the hypothesis on whether polarization has increased over the years on Twitter in Publication IV. We test this hypothesis along the various dimensions on Twitter: retweets, follow and content. This is the first long-term analysis of polarization on Twitter.

We used a dataset focused on a set of public seed Twitter accounts: politicians and media outlets, with known political leaning. From these seed users we then crawl outwards by collecting data for users who follow or retweet the seed users. Details as follows.

Seed Accounts. Our point of departure is a list with two types of polarized seed accounts. The first type consists of presidential/vice presidential candidates and their parties (see the political accounts in Table 5.1) for the last eight years. The second type consists of popular media accounts listed in Table 5.1. The list of media outlets was obtained from a report by the Pew Research Center on polarization and media habits.[1]

Following Users. For each seed user, we obtained all their followers. The combined set of all followers for all seed accounts gave us a total of 140M users. We estimated the time when a user followed a particular seed account using the method proposed by Meeder et al. [95]. This method is based on the fact that the Twitter API returns followers in the reverse chronological order in which they followed and we can lower bound the follow time using the account creation date of a user. So, as at least some of @BarackObama's followers started to follow him right after creating their Twitter account, this leads to temporal bounds for the other followers as well. These estimates are reported to be fairly accurate when

[1] http://www.journalism.org/2014/10/21/political-polarisation-media-habits/

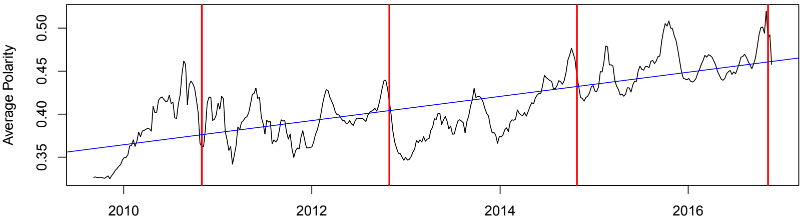


Figure 5.5. Content polarization over time. Red vertical lines indicate the mid term and main elections in the U.S. The blue line is the linear fit, which has a non-zero slope tested using a t-test (p<0.0001).

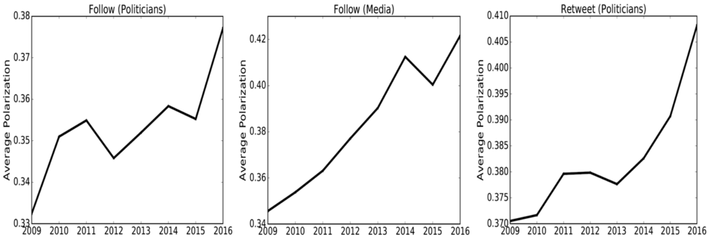


Figure 5.6. Follow (left, middle) and retweet (right) effects over time for politicians (left, right) and media (middle) seed accounts.

estimating follow times for users with millions of followers. For our analysis, we used all cases with estimated follow dates from January 2009 onwards.

Retweeting Users. For the set of seed politicians, we obtained all their public, historic tweets.[2] The earliest tweets in this collection date back to 2006. For each collected tweet, we used the Twitter API to collect up to 100 retweets. This gave us a set of 1.3M unique users who retweeted a political entity since 2006. We randomly sampled 50% of these users (679,000), and used the Twitter API to get 3,200 of their most recent tweets in December 2016. This gave us around 2.5 billion tweets. Though we have tweets dating back to 2007, we only consider tweets from September 2009 onwards in the analysis since the volume for earlier tweets is low.

Based on this dataset, we find that polarization on Twitter has increased over the last 8 years, in terms of various dimensions such as following, retweeting and content produced. Figure 5.5 shows polarization of content over time, as measured by a measure of hashtag polarization proposed by Weber et al [127].

This trend is consistent across measures, and depending on the measure (follow, retweet or content), the relative change is 10%-20% (e.g. see Figure 5.6 for results using follow and retweet measures). Our study is one of very few with such a long-term perspective, encompassing two U.S. presidential elections and two mid-term elections, providing a rare longitudinal analysis. For more details on the measures and the dataset, please refer to Publication IV.

[2] Since the Twitter API restricts us to the last 3200 tweets, we used a public tool to get all historic tweets `https://github.com/Jefferson-Henrique/GetOldTweets-python`

6. Reducing Polarization

In the previous sections, we identified polarized topics and understood some of their properties over time. We also show that polarization on Twitter has been increasing over the last decade. In this section, we devise algorithmic solutions to handle this increasing polarization. In particular, we propose two methods.

First, we propose reducing polarization by connecting Twitter users with opposing viewpoints. This is based on the idea of people being stuck in echo chambers, where they only see content from their own side and are not exposed and hence are not aware of any content from the other side. Next, we take an information propagation approach and propose an idea to spread information in the network in such a way that every user gets a balanced access to information from both sides of the debate.

6.1 Connecting Opposing Views

Society is often polarized by controversial issues that split the population into groups with opposing views. When such issues emerge on social media, we often observe the creation of "echo chambers", i.e., situations where like-minded people reinforce each other's opinion, but do not get exposed to the views of the opposing side.

In this section, we study an algorithmic technique for bridging these chambers, and thus reduce polarization. Usually, discussions on polarized topics involve a fair share of "retweeting" or "sharing" opinions of authoritative figures that the user agrees with. Therefore, it is natural to model the discussion as an *endorsement graph* or a retweet graph: a vertex v represents a user, and a directed edge (u,v) represents the fact that user u endorses the opinion of user v.

We then cast our problem as an *edge-recommendation problem* on this graph. The goal of the recommendation is to reduce the *controversy score* of the graph (RWC), which is measured by a metric based on random walks (see Section 4.1.1 for details). In particular, given a metric that measures how polarized an issue is on social media (RWC), or how biased is a user who discusses the issue

(RWC_{user}), our goal is to find a small number of edges, called *bridges*, which minimize these measures (RWC or RWC_{user}). That is, we seek to propose (content produced by) a user v to another user u, aiming that u endorses v by spreading their opinion. This action would create a new edge (a bridge) in the endorsement graph, thus reducing the polarization score of the graph (topic) or the user itself.

Clearly, some bridges are more likely to materialize than others. For instance, people in the "center" might be easier to convince than people on the two extreme ends of the political spectrum [87]. We take this issue into account by modeling an *acceptance probability* for a bridge as a separate component of the model. This component can be implemented by any generic link-prediction algorithm that gives a probability of materialization to each non-existing edge. However, we propose a simple model based on RWC_{user}(detailed in Section 4.1.2) [55], which captures the dynamics and properties of the endorsement graph. Therefore, we seek bridges that minimize the *expected* controversy score, according to their acceptance probabilities.

We consider two variants of the problem. First, a global version where we aim to find the best possible connections to make for the good of the entire society [56], and second, a more practical version which deals with individual level, i.e., propose the best recommendations for a user that will reduce her polarization [54].

We can define the first variant of our problem formally as follows:

Problem 1 (k-EDGEADDITION). *Given a graph $G(V,E)$ whose vertices are partitioned into two disjoint sets X and Y ($X \cup Y = V$ and $X \cap Y = \emptyset$), and an integer k, find a set of k edges $E' \subseteq V \times V \setminus E$ to add to G and obtain a new graph $G' = (V, E \cup E')$, so that the controversy score* $\mathrm{RWC}(G', X, Y)$ *is minimized.*

6.1.1 Acceptance probability

Problem 1 seeks the edges that lead to the lowest RWC score *if added* to the graph. In a recommendation setting, however, the selected edges do not always materialize (e.g., the recommendation might be rejected by the user). In such a setting, it is more appropriate to consider edges that minimize the RWC score *in expectation*, under a probabilistic model $\mathbb{A}$ that provides the probability that a set of edges are accepted once recommended. This consideration leads us to the following formulation of our problem.

Problem 2 (k-EDGEADDITIONEXPECTATION). *Given a graph $G = (V,E)$ whose vertices are partitioned into two disjoint sets X and Y ($X \cup Y = V$ and $X \cap Y = \emptyset$), and an integer k, find a set of k edges $E' \subseteq V \times V \setminus E$ to add to G and obtain a new graph $G' = (V, E \cup E')$, so that the expected controversy score E_A*[$\mathrm{RWC}(G', X, Y)$] *is minimized under acceptance model $\mathbb{A}$.*

We build such an acceptance model $\mathbb{A}$ on the feature of *user polarity* described in Section 4.1.2. We employ user polarity as a feature for our acceptance model

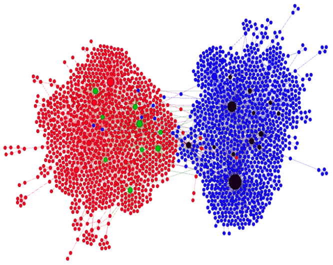


Figure 6.1. An example retweet graph for the topic #russia_march. The green and black dots indicate nodes picked by our algorithm.

because, intuitively, we expect users from each side to accept content from different sides with different probabilities, and we assume these probabilities are encoded in, and can be learned from, the graph structure itself. For example, a user with polarity close to -1 is more likely to endorse a user with a negative polarity than a user with polarity $+1$.

Now let u and v be two users with polarity R_u and R_v, respectively. Moreover, assume that u is not connected to v in the current instantiation of the graph. Let $p(u,v)$ be the probability that u accepts a recommendation to connect to v. We estimate $p(u,v)$ from training data. Given a dataset of user interactions, we estimate $p(u,v)$ as the fraction

$$N_e(R_u,R_v)/N_x(R_u,R_v)$$

where $N_x(R_u,R_v)$ and $N_e(R_u,R_v)$ are the number of times a user with polarity R_v was *exposed to* or *endorsed* (respectively) content generated by a user of polarity R_u. $N_x(R_u,R_v)$ is computed by assuming that if v follows u, v is exposed to all content generated by u.

Figure 6.1 shows an example retweet network with edges recommended by our algorithm added.

6.1.2 User level recommendation

Problems 1 and 2 solve the problem of finding the best pairs of users to connect in a network. These are the best pairs in an ideal situation that help to make the entire society (or the topic) less polarized. However, even with the addition of the acceptance probabilities, there is no guarantee that these pairs of users will accept a recommendation to connect.

A simpler, more realistic variant of the above problem is to make connections at a user level, that is, to help a user reduce their polarization. Based on this, we define the following problem.

Problem 3 (k-EDGEADDITIONUSER). *Given a graph $G(V,E)$, a user u and an* *integer k, find a set of k edges $E' \subseteq V \times V \setminus E$ to add to G from u so that the controversy score* $\mathrm{RWC}_{user}(u)$ *is minimized.*

Problem 3 is much more practical and feasible in real world. Though our main focus is to connect users with content that expresses a contrarian point of view, we also want to maximize the chances of such a recommendation being endorsed by the user. As we propose above, taking into account the acceptance probability is one way to address this issue. We can also take into account other factors such as:

Topic diversity. We want to ensure that the recommendations made for a user are topically diverse and similar to the interests of the user. To achieve this, for each user, we compute a vector t_u that contains the topics extracted from the tweets written and the items shared by the user. Similarly, we extract a vector of topics t_i for each content item being recommended. Topics are defined as a *named entity*, and we extract them using the tool TagMe.[1] Given a user vector t_u, we compute the cosine similarity with all item vectors t_i, and rank items in a decreasing order of cosine similarity.

Popularity on either side. We can also take into account the popularity of the recommended items, so that users receive content that is popular and, likely, of good quality. For each item, we compute a popularity score as the maximum number of retweets obtained by a tweet that contains this item.

Given these different factors, we can produce a final recommendation for the user by simply modeling the recommendation problem as a weighted rank aggregation problem.

To evaluate the recommendations generated by our algorithm, we run an online user study involving around 7,000 Twitter users who were active participants on the 2016 U.S. election result night. For each user in the study, we generate two recommended items that are personalized based on their Twitter activity: one item is highly contrarian, while the other is more likely to be accepted, according to our model. Our expectation is that users enjoy reading the item with high acceptance probability, and disagree with the contrarian item. Each user was contacted using a Twitter bot that sent automated messages. We find that most users indeed enjoy reading the item with high acceptance, and disagree with the contrarian item. Details of the algorithms and experiments can be found in Publication III and Publication VI.

6.2 Spreading information

In the previous section, we looked at methods to reduce polarization by convincing people to connect to others with an opposing viewpoint. In this section, we take a look at the problem of reducing polarization from a different perspective, instead looking at spreading information that balances users exposure to news.

[1] `https://services.d4science.org/web/tagme`

We consider social-media discussions around a topic that are characterized by two conflicting viewpoints. Let us refer to these viewpoints as *campaigns*. Our approach follows the popular paradigm of influence maximization [76]: we want to select a small number of seed users for each campaign so as to maximize the number of users who are *exposed to both campaigns*. In contrast to existing work on competitive viral marketing, we do not consider the problem of finding an optimal *selfish strategy* for each campaign separately. Instead we consider a *certalized agent* responsible for balancing information exposure for the two campaings.

Consider the following motivating examples on how such an approach could reduce polarization.

Example 1: Prominent social-media companies, like Facebook and Twitter, have been called to act as arbiters so as to prevent ideological isolation and polarization in the society. The motivation for companies to assume this role could be for improving their public image or due to government policies.[2] Consider a controversial topic being discussed in social-media platform X, which has led to polarization. Platform X has the ability to algorithmically detect polarization [52], identify the influential users on each side, and estimate the influence among users [42, 61]. As part of a new polarization reduction service, platform X would like to disseminate two high-quality and thought-provoking dueling op-eds, articles, one for each side, that present the arguments of the other side in a fair manner. Assume that X is interested in following a viral-marketing approach. Which users should X target, for each of the two articles, so that people in the network are informed in the most balanced way?

Example 2: Government organization Y is initiating a program to help assimilate foreigners who have newly arrived in the country. Part of the initiative focuses on bringing the communities of foreigners and locals closer in social media. Organization Y is interested in identifying individuals who can help spreading news of one community into the other.

From a technical standpoint, we consider the following problem setting: We assume that information is propagated in the network according to the *independent-cascade model* [76]. We assume that there are two opposing campaigns, and for each one there is a set of initial seed nodes, I_1 and I_2, which are not necessarily distinct. Furthermore, we assume that the users in the network are exposed to information about campaign i via diffusion from the set of seed nodes I_i. The diffusion in the network may occur with independent or correlated probabilities for the two campaigns; we consider both settings to which we are referring as *heterogeneous* or *correlated*.

The objective is to recruit two additional sets of seed nodes, S_1 and S_2, for the two campaigns, with $|S_1|+|S_2| \leq k$, for a given budget k, so as to maximize the expected number of balanced users, i.e., the users who are exposed to information from both campaigns (or from none!).

Although our approach is inspired by the large body of work on information

[2] For instance, Germany is now fining Facebook for the spread of fake news.

propagation, and resembles previous problem formulations for competitive viral marketing, there are significant differences and novelties. In particular:

- This is the first paper to address the problem of *balancing information exposure* to reduce polarization, using the information-propagation methodology.

- The objective function that best suits our problem setting is related to the *size of the symmetric difference* of users exposed to the two campaigns. This is in contrast to previous settings that consider functions related to the *size of the coverage* of the campaigns.

- As a technical consequence of the previous point, our objective function is neither *monotone* nor *submodular* making our problem more challenging. Yet we are able to analyze the problem structure and provide algorithms with approximation guarantees.

- While most previous papers consider selfish agents, and provide bounds on *best-response* strategies (i.e., move of the last player), we consider a centralized setting and provide bounds for a global objective function.

We start with a directed graph $G = (V, E, p_1, p_2)$ representing a social network. We assume that there are two distinct campaigns that propagate through the network. Each edge $e = (u, v) \in E$ is assigned two probabilities, $p_1(e)$ and $p_2(e)$, representing the probability that a post from vertex u will propagate (e.g., it will be reposted) to vertex v in the respective campaigns. Given a seed set S, we write $r_1(S)$ and $r_2(S)$ for the vertices that are reached from S using the independent cascade model.

Given a directed graph, initial seed sets for both campaigns and a budget, we ask to find additional seeds that would balance the information adopted by the vertices. More formally:

Problem 4 (BALANCE)**.** *Let $G = (V, E, p_1, p_2)$ be a directed graph, and two sets I_1 and I_2 of initial seeds of the two campaigns. Assume that we are given a budget k. Find two sets S_1 and S_2, where $|S_1| + |S_2| \leq k$ maximizing*

$$\Phi(S_1, S_2) = E\left[|V \setminus (r_1(I_1 \cup S_1) \bigtriangleup r_2(I_2 \cup S_2))|\right].$$

The objective function $\Phi(S_1, S_2)$ is the expected number of vertices that are either reached by both campaigns or remain oblivious to both campaigns.

In Publication IX, we show that it is **NP**-hard. We develop different algorithms by decomposing the above objective function $\Phi(S_1, S_2)$, one of which has a $(1 - 1/e)/2$ approximation guarantee.

We experimentally evaluate our methods, on several real-world (and realistic) datasets, collected from Twitter, for different polarized topics. Details of our algorithms and experiments can be found in Publication IX.

7. Limitations and future work

In this section, we discuss some of the limitations of the approaches presented in this thesis. Next, we try to provide some avenues for improvement and questions to ponder over.

7.1 Limitations

The thesis studies polarization, a timely and relevant topic, which spans a wide range of scientific disciplines, including social science, political science, psychology, design and computer science. In our study, we make some assumptions and simplifications to make the thesis tractable. In this section, we describe a few limitations of our work and suggest potential steps to alleviate these limitations, wherever possible.

Twitter only. The thesis is based primarily on Twitter-specific details and all experiments are done using Twitter. An important consequence of depending entirely on Twitter is the question of how generalizable our approaches are. Twitter has only a certain reach — according to a recent Pew research center survey [110], only 18% of the U.S. adults use Twitter as a source of information. It also has its own biases in terms of demographics [18], e.g., users over the age of 65 might not be well represented.

While this is certainly a limitation, Twitter is one of the main venues for online public discussion, and one of the few for which data is available. Hence, Twitter is a natural choice. In addition, our methods generalize well to datasets from other social media and the Web.

Choice of data. In many of our experiments, we manually pick the polarized topics, defined as hashtags or a group of hashtags, which might be limiting and introduce bias. These hashtags are picked from common knowledge (e.g., say, assuming that #obamacare is a polarized topic). Since there is no clear way to evaluate whether this is true in all cases, this might be a limitation.

To counter issues with specificity of defining the topics, we select topics that represent a broad set of typical polarized issues coming from religious, societal, racial, and political domains. Unfortunately, ground truths for polarized topics are hard to find, especially for ephemeral issues. Moreover, the hashtags represent the intuitive notion of polarization that we strive to capture, so human judgment is an important ingredient we want to use.

Contradictory results. An important consequence of the above limitation is clearly evident in producing slightly contradictory results in this thesis. For instance, in Publication V, we find that there is no consistent trend in polarization for the long term polarized topics obamacare, abortion, guncontrol and fracking, where as in Publication IV, we find a consistent increase in long term polarization. This is because we use different datasets — in Publication V we consider specific topics and collect data pertaining to those topics, while in Publication IV we collect a larger dataset.

Since we do not have access to the complete data and we work with subsets, we can try to be as thorough as possible and not introduce biases in our measurements, but this is not always possible. As we mentioned in Chapter 2, there is no consensus in the literature on many of the topics we study, including, whether polarization in the society is increasing, and whether social media helps creating echo chambers.

Only two sides. In the thesis, we make a strong assumption that all polarized topics we deal with have two clear opposing sides and that these two sides can be obtained by clustering the retweet graph. Not all polarized discussions involve only two sides with opposing views. Oftentimes discussions are multifaceted, and there are three or more competing views on the field. Although this is a big assumption, it makes the analysis and development of algorithms easier. The principles behind our methods neatly generalize to multi-sided polarized topics. We acknowledge that there is a need to develop techniques that do not strictly depend on this assumption and defer such cases for future work.

No use of Content. Our methods are primarily network based. We (mostly) do not make use of the language of the tweets. This is a deliberate choice that allows us to deal with multiple topics from different domains and languages. However, depending solely on network features hurts our case because we miss important signals from text. For instance, in the conference version of our work [52], as well as in Publication VIII and Publication VII we show that using text based-features (e.g., sentiment) also helps in identifying polarized topics.

Focus on algorithms. One of the main challenges we deal with in this thesis is to develop algorithms to reduce polarization. While developing and performing experiments in Publication VI, we observe that reducing polarization is not just a technical problem, but also a social and psychological problem. We should also take into account the psychological and social aspects and literature into account when considering recommending content that goes against a user's viewpoint, to avoid pitfalls such as the Backfire effect [114]. To a certain extent we are already trying to address these issues (e.g., using the acceptance probability), but ideally one should work more closely with experts from other fields.

That said, it is still very important to work on computational tools, in tandem with developments in other fields, like social science and psychology to help us scale the findings from those fields to a large number of users and help be deployed to millions of users. We feel that our results contribute to this direction.

Real-world evaluation. Related to the above point, another limitation of the thesis is that though we propose techniques to reduce polarization, there is no way to objectively evaluate if our methods actually work. An objective evaluation is only possible with access to click/browse logs, which are unfortunately only accessible from within the social media companies (like Facebook or Twitter). We try our best to elicit response from real social network users, by creating a bot and sending out messages to users (Publication VI), but still, this is limited.

Combining different signals. In Chapter 4, we present various methods using different types of interactions to decide whether a topic is polarized or not. We make the assumption that certain patterns in these interaction networks, e.g. clustered structure of the retweet network, help us identify polarized topics. We must note though, that this is just a necessary condition but not a sufficient condition. A topic might have a clustered retweet structure but not be polarized, e.g. a retweet graph for a promotion campaigns by different organizations might also have a clustered structure. One way to tackle this issue is to combine the different signals we propose in order to make a conclusion. Using multiple approaches together (e.g. retweet network, reply network and content), we can be more confident that the topic is polarized.

7.2 Future work

In this section, we discuss some directions for future work.

Modeling. One of the important areas in the study of polarization deals with building generative models to explore opinion formation and the dynamics of polarization. In this thesis, we do not explore this direction at all. However, we can build upon the findings from the thesis to understand and build better models for polarization. Our observations pave the way to the development of models for evolution of interaction networks in polarized topics, similar to how studies about measuring the web and social media were the stepping-stone to developing models for them. We can also design probabilistic generative models to capture the observed effects of polarization in terms of content and network features. Our findings (in Publication X) show the interaction between network importance and the content produced and consumed by a user. Most of the existing models for dynamics of opinion formation and polarization on social networks either use exclusively content features, or use a dynamic process on a fixed random network [12]. However, in light of our results, a comprehensive model for polarization should affect not only the opinion spread over the social network, but also the structure of the network itself.

Content. As mentioned in Section 7.1, most of our methods do not take content into account. In our brief experiments with content, we find that though using content might have its own limitations, it carries some signal of value and helps in better identifying polarization. In the future, we could look at two aspects of content analysis. First, using existing tools like topic modeling, we could extract topics from text to define at a better granularity what users might be interested in. For instance, consider again the topic #obamacare. A user might be supporting of #obamacare in general, but might be opposed to a special subtopic, say, Hillary Clinton's plan to reform #obamacare. To find such granular details, we need to analyze the content. Second, the interaction between content and network in the context of polarization has also not been explored well. These are interesting directions to explore.

Generalization. Most of our methods are defined in a Twitter-specific terminology, but almost all other social networks have equivalent mechanisms, e.g., retweet on Twitter has a similar meaning as "share" on Facebook, or "reblog" on Tumblr. Though these methods generalize across social networks, the results of applying these methods might vary because of different populations [18], or differences in the intended usage of other social-media platforms. It will be interesting to see to what extent these methods generalize to other social networks.

Tackling root causes. Assuming that excessive polarization is bad for the society, what can we do to handle the root cause of it? As we saw in Section 2.1, one of the root causes of polarization is the various types of biases that exist at various levels in the society. As we see in Figure 2.1, user biases constitute a major portion in this. Though it is not easy or practical to tackle and find solutions to all these, we could start with imbibing simple traits such as valuing opinions from the other side and building tools that engage users in a healthy discussion. One way to achieve such an objective would be to have applications for serving a "healthier" and more balanced news diet to social-media users [79].

Ethical questions. Finally, this thesis touches upon topics that raise certain ethical questions. What does it mean to depolarize the society? Polarization by itself may not be a totally negative phenomenon. Several studies [101, 35] argue that some level of polarization is needed for a democracy. Mutz el al. [101] state that *"a democracy needs deliberation, and polarization enable such a deliberation to happen in the public, to a certain extent, thus informing people about the issues and arguments from different sides"*. Given such a setting, it is of paramount importance to understand how we can create constructive polarization. But how can we decide what constitutes constructive polarization that needs to be encouraged? As we design algorithms that make recommendations to people, how can ensure that these recommendations are right for people? Does it constitute to manipulating their decision making? It is important to ponder answers to these questions as computer scientists before developing such systems. There is a recent interest in the field of transparency and explainability of algorithms [45], which might help answering some of these questions.

8. Conclusions

As social media is coming of age and soon teenagers will no longer remember a pre-social-media era, we need to be aware of both the positive and negative effects that come in reinventing how users get their information. Though the internet and social media have been envisioned as places that create an open forum for discussion, they also create avenues for isolation and hatred towards other users who do not necessarily agree with your opinion.

We could clearly observe the influence of nefarious actors such as automated bots, fake news and propaganda in the outcomes of recent high profile events such as the 2016 U.S. presidential elections & brexit, and the potential role of polarization in abetting these actors. Thus, understanding polarization and aspects that surround it are the need of the hour.

This thesis contributes in improving the understanding of polarization, mostly from a computer-science perspective. We provide automated tools to identify polarization on social media and use these tools to study properties of polarized topics over time. We then develop algorithms to reduce polarization by connecting users with opposing viewpoints and by prompting information to spread in a network in such a way that existing viewpoints received a balanced coverage.

As we hint in Chapter 7, these are not just problems in computer science. Our thesis just scratches the surface in getting a better understanding of polarization. In our work, we highlight some potential techniques to understand and handle polarization and placed the results within the context of a larger debate. A close collaboration of different fields, including input from humanities and computer science are needed to completely tackle the issue of polarization.

References

[1] Alan I Abramowitz and Kyle L Saunders. Is polarization a myth? *The Journal of Politics*, 70(2):542–555, 2008.

[2] Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *LinkKDD*, pages 36–43, 2005.

[3] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th International World Wide Web Conference*, pages 835–844. ACM, 2007.

[4] Leman Akoglu. Quantifying political polarity based on bipartite opinion networks. In *Proceedings of the 8th AAAI International Conference on Web and Social Media*, 2014.

[5] Eduardo Alemán, Ernesto Calvo, Mark P Jones, and Noah Kaplan. Comparing cosponsorship and roll-call ideal points. *Legislative Studies Quarterly*, 34(1):87–116, 2009.

[6] Kelsey Allen, Giuseppe Carenini, and Raymond T Ng. Detecting disagreement in conversations using pseudo-monologic rhetorical structure. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 1169–1180, 2014.

[7] Noga Alon, Michal Feldman, Ariel D Procaccia, and Moshe Tennenholtz. A note on competitive diffusion through social networks. *IPL*, 110(6):221–225, 2010.

[8] Md Tanvir Al Amin, Charu Aggarwal, Shuochao Yao, Tarek Abdelzaher, and Lance Kaplan. Unveiling polarization in social networks: A matrix factorization approach. Technical report, IEEE, 2017.

[9] Jisun An, Daniele Quercia, and Jon Crowcroft. Partisan sharing: facebook evidence and societal consequences. In *COSN*, pages 13–24, 2014.

[10] Clio Andris, David Lee, Marcus J Hamilton, Mauro Martino, Christian E Gunning, and John Armistead Selden. The rise of partisanship and super-cooperators in the us house of representatives. *PloS one*, 10(4):e0123507, 2015.

[11] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, 2015.

[12] Sven Banisch and Eckehard Olbrich. Opinion polarization by learning from social feedback. *arXiv preprint arXiv:1704.02890*, 2017.

[13] Pablo Barberá. How social media reduces mass political polarization. evidence from germany, spain, and the us. *Job Market Paper, New York University*, 46, 2014.

[14] Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91, 2015.

[15] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.

[16] Alessandro Bessi, Guido Caldarelli, Michela Del Vicario, Antonio Scala, and Walter Quattrociocchi. Social Determinants of Content Selection in the Age of (Mis)Information. In *Social Informatics*, pages 259–268, 2014.

[17] Shishir Bharathi, David Kempe, and Mahyar Salek. Competitive influence maximization in social networks. In *WINE*, 2007.

[18] Grant Blank. The digital divide among twitter users and its implications for social research. *Social Science Computer Review*, 2016.

[19] Robert Bond and Solomon Messing. Quantifying social media's political space: Estimating ideology from publicly revealed preferences on facebook. *American Political Science Review*, 109(1):62–78, 2015.

[20] Adam Bonica. Ideology and interests in the political marketplace. *American Journal of Political Science*, 57, 2013.

[21] Allan Borodin, Mark Braverman, Brendan Lucier, and Joel Oren. Strategyproof mechanisms for competitive influence in networks. In *Proceedings of the 22nd International World Wide Web Conference*, pages 141–150, 2013.

[22] Shelley Boulianne. Social media use and participation: A meta-analysis of current research. *Information, Communication & Society*, 18(5):524–538, 2015.

[23] Levi Boxell, Matthew Gentzkow, and Jesse M Shapiro. Greater internet use is not associated with faster growth in political polarization among us demographic groups. *Proceedings of the National Academy of Sciences*, page 201706588, 2017.

[24] Aaron Bramson, Patrick Grim, Daniel J Singer, Steven Fisher, William Berger, Graham Sack, and Carissa Flocken. Disambiguation of social polarization concepts and measures. *The Journal of Mathematical Sociology*, 40(2):80–111, 2016.

[25] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th International World Wide Web Conference*, pages 665–674, 2011.

[26] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th International World Wide Web Conference*, pages 675–684. International World Wide Web Conferences Steering Committee, 2011.

[27] Zoey Chen and Jonah Berger. When, why, and how controversy causes conversation. *Journal of Consumer Research*, 40(3):580–593, 2013.

[28] Yoonjung Choi, Yuchul Jung, and Sung-Hyon Myaeng. Identifying controversial issues and their sub-topics in news articles. In *Pacific-Asia Workshop on Intelligence and Security Informatics*, pages 140–153. Springer, 2010.

[29] Peter Cogan, Matthew Andrews, Milan Bradonjic, W Sean Kennedy, Alessandra Sala, and Gabriel Tucci. Reconstruction and analysis of twitter conversation graphs. In *First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, pages 25–31. ACM, 2012.

[30] Raviv Cohen and Derek Ruths. Classifying political orientation on twitter: It's not easy! In *Proceedings of the 7th AAAI International Conference on Web and Social Media*, 2013.

[31] Mauro Coletto, Kiran Garimella, Aristides Gionis, and Claudio Lucchese. A motif-based approach for identifying controversy. In *Proceedings of the 11th AAAI International Conference on Web and Social Media*. AAAI, 2017.

[32] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, 64(2):317–332, 2014.

[33] Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. *Proceedings of the 5th AAAI International Conference on Web and Social Media*, 133:89–96, 2011.

[34] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *Proceedings of the 3rd Inernational Conference on Social Computing (SocialCom)*, pages 192–199. IEEE, 2011.

[35] Lincoln Dahlberg. Rethinking the fragmentation of the cyberpublic: from consensus to contestation. *New media & society*, 9(5):827–847, 2007.

[36] Manlio De Domenico, Antonio Lima, Paul Mougel, and Mirco Musolesi. The anatomy of a scientific rumor. *Scientific reports*, 3, 2013.

[37] Stefano DellaVigna and Ethan Kaplan. The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234, 2007.

[38] Nicholas DiFonzo, Jerry Suls, Jason W Beckstead, Martin J Bourgeois, Christopher M Homan, Samuel Brougher, Andrew J Younge, and Nicholas Terpstra-Schwab. Network structure moderates intergroup differentiation of stereotyped rumors. *Social Cognition*, 32(5):409, 2014.

[39] Kenneth L Dion. Cohesiveness as a determinant of ingroup-outgroup bias. *Journal of Personality and Social Psychology*, 28(2):163, 1973.

[40] Shiri Dori-Hacohen and James Allan. Automated controversy detection on the web. In *European Conference on Information Retrieval*, pages 423–434. Springer, 2015.

[41] Sergei N Dorogovtsev and José FF Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. OUP Oxford, 2013.

[42] Nan Du, Le Song, Manuel Gomez Rodriguez, and Hongyuan Zha. Scalable influence estimation in continuous-time diffusion networks. In *Advances in neural information processing systems*, pages 3147–3155, 2013.

[43] Arthur Edwards. (how) do participants in online discussion forums create ‘echo chambers’?: The inclusion and exclusion of dissenting voices in an online forum about climate change. *Journal of Argumentation in Context*, 2(1):127–150, 2013.

[44] Joan-Maria Esteban and Debraj Ray. On the measurement of polarization. *Econometrica: Journal of the Econometric Society*, pages 819–851, 1994.

[45] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[46] Leon Festinger. *A theory of cognitive dissonance*, volume 2. Stanford university press, 1962.

[47] Morris P Fiorina and Samuel J Abrams. Political polarization in the american public. *Annual Review of Political Science*, 11:563–588, 2008.

[48] Peter Fischer, Dieter Frey, Claudia Peus, and Andreas Kastenmüller. The theory of cognitive dissonance: State of the science and directions for future research. In *Clashes of Knowledge*, pages 189–198. Springer, 2008.

[49] Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, 2016.

[50] Jonathan L Freedman and David O Sears. Selective exposure. *Advances in experimental social psychology*, 2:57–97, 1965.

[51] Dieter Frey. Recent research on selective exposure to information. *Advances in experimental social psychology*, 19:41–80, 1986.

[52] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy in social media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 33–42. ACM, 2016.

[53] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. The ebb and flow of controversial debates on social media. In *Proceedings of the 11th AAAI International Conference on Web and Social Media*. AAAI, 2017.

[54] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Factors in recommending contrarian content on social media. In *Proceedings of the 10th Annual ACM Web Science Conference*. ACM, 2017.

[55] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy in social media. In *Transactions on Social Computing*. ACM, 2017.

[56] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Reducing controversy by connecting opposing views. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*. ACM, 2017.

[57] R Kelly Garrett. Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication*, 14(2):265–285, 2009.

[58] Matthew Gentzkow and Jesse M Shapiro. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839, 2011.

[59] Eric Gilbert, Tony Bergstrom, and Karrie Karahalios. Blogs are echo chambers: Blogs are echo chambers. In *42nd Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2009.

[60] CW Gini. Variability and mutability, contribution to the study of statistical distribution and relaitons. *Studi Economico-Giuricici della R*, 1912.

[61] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028. ACM, 2010.

[62] Sanjeev Goyal, Hoda Heidari, and Michael Kearns. Competitive contagion in networks. *Games and Economic Behavior*, 2014.

[63] Eduardo Graells-Garrido, Mounia Lalmas, and Daniele Quercia. Data portraits: Connecting people of opposing views. *arXiv preprint arXiv:1311.4658*, 2013.

[64] Eduardo Graells-Garrido, Mounia Lalmas, and Daniele Quercia. People of opposing views can share common interests. In *Proceedings of the 23rd International World Wide Web Conference Companion*, pages 281–282. International World Wide Web Conferences Steering Committee, 2014.

[65] Catherine Grevet, Loren G Terveen, and Eric Gilbert. Managing political differences in social media. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1400–1408. ACM, 2014.

[66] Max Grömping. 'Echo Chambers' Partisan Facebook Groups during the 2014 Thai Election. *Asia Pacific Media Educator*, 24(1):39–59, 2014.

[67] Pedro Henrique Calais Guerra, Wagner Meira Jr, Claire Cardie, and Robert Kleinberg. A measure of polarization on social media networks based on community boundaries. In *Proceedings of the 7th AAAI International Conference on Web and Social Media*. AAAI, 2013.

[68] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(2):17–28, 2013.

[69] Martin Harrison. *TV News, Whose Bias?: A Casebook Analysis of Strikes, Television and Media Studies*. Hermitage [England]: Policy Journals, 1985.

[70] Kyle A Heatherly, Yanqin Lu, and Jae Kook Lee. Filtering out the other side? cross-cutting and like-minded discussions on social networking sites. *New media & Society*, 19(8):1271–1289, 2017.

[71] Alfred Hermida, Fred Fletcher, Darryl Korrell, and Donna Logan. Your friend as editor: the shift to the personalized social news stream. In *Future of Journalism Conference*, pages 8–9, 2011.

[72] Marc J Hetherington. Resurgent mass partisanship: The role of elite polarization. *American Political Science Review*, 95(3):619–631, 2001.

[73] Myungha Jang, John Foley, Shiri Dori-Hacohen, and James Allan. Probabilistic approaches to controversy detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2069–2072. ACM, 2016.

[74] Denise B Kandel. Homophily, selection, and socialization in adolescent friendships. *American journal of Sociology*, 84(2):427–436, 1978.

[75] George Karypis and Vipin Kumar. METIS - Unstructured Graph Partitioning and Sparse Matrix Ordering System, 1995.

[76] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[77] Manfred Klenner, Michael Amsler, Nora Hollenstein, and Gertrud Faaß. Verb polarity frames: a new resource and its application in target-specific polarity classification. In *KONVENS*, pages 106–115, 2014.

[78] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 417–432. ACM, 2017.

[79] Juhi Kulshrestha, Muhammad Bilal Zafar, Lisette Espin Noboa, Krishna P Gummadi, and Saptarshi Ghosh. Characterizing information diets of social media users. In *Proceedings of the 9th AAAI International Conference on Web and Social Media*. AAAI, 2015.

[80] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[81] Janette Lehmann, Bruno Gonçalves, José J. Ramasco, and Ciro Cattuto. Dynamical Classes of Collective Attention in Twitter. In *Proceedings of the 21st International World Wide Web Conference*, pages 251–260. ACM, 2012.

[82] Yphtach Lelkes. Mass polarization: Manifestations and measurements. *Public Opinion Quarterly*, 80(S1):392–410, 2016.

[83] Yphtach Lelkes, Gaurav Sood, and Shanto Iyengar. The hostile audience: The effect of access to broadband internet on partisan affect. *American Journal of Political Science*, 61(1):5–20, 2017.

[84] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470. ACM, 2008.

[85] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 177–187. ACM, 2005.

[86] Q Vera Liao and Wai-Tat Fu. Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2359–2368. ACM, 2013.

[87] Q Vera Liao and Wai-Tat Fu. Can you hear me now?: mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 184–196. ACM, 2014.

[88] Q Vera Liao and Wai-Tat Fu. Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2745–2754. ACM, 2014.

[89] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[90] Zhe Liu and Ingmar Weber. Is Twitter a public sphere for online conflicts? A cross-ideological and cross-hierarchical look. In *SocInfo*, pages 336–347. Springer, 2014.

[91] Charles G Lord, Lee Ross, and Mark R Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.

[92] Haokai Lu, James Caverlee, and Wei Niu. Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 213–222. ACM, 2015.

[93] Antonis Matakos, Evimaria Terzi, and Panayiotis Tsaparas. Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery*, 31(5):1480–1505, 2017.

[94] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[95] Brendan Meeder et al. We know who you followed last summer: inferring social link creation times in twitter. In *Proceedings of the 20th International World Wide Web Conference*, pages 517–526, 2011.

[96] Yelena Mejova, Amy X Zhang, Nicholas Diakopoulos, and Carlos Castillo. Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152*, 2014.

[97] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. Political polarization & media habits. *http://www.journalism.org/2014/10/21/political-polarization-media-habits*, 2014.

[98] AJ Morales, Javier Borondo, Juan Carlos Losada, and Rosa M Benito. Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3):033114, 2015.

[99] Sean A Munson, Stephanie Y Lee, and Paul Resnick. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of the 7th AAAI International Conference on Web and Social Media*, 2013.

[100] Sean A Munson and Paul Resnick. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1457–1466. ACM, 2010.

[101] Diana C Mutz. The consequences of cross-cutting networks for political participation. *American Journal of Political Science*, pages 838–855, 2002.

[102] Seth A Myers and Jure Leskovec. Clash of the contagions: Cooperation and competition in information diffusion. In *Proceedings of the 14th International Conference on Data Mining*, pages 539–548, 2012.

[103] Mark Newman, Albert-Laszlo Barabasi, and Duncan J Watts. *The structure and dynamics of networks*. Princeton University Press, 2011.

[104] Nam P Nguyen, Guanhua Yan, My T Thai, and Stephan Eidenbenz. Containment of misinformation spread in online social networks. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 213–222, 2012.

[105] Ryosuke Nishi, Taro Takaguchi, Keigo Oka, Takanori Maehara, Masashi Toyoda, Ken-ichi Kawarabayashi, and Naoki Masuda. Reply trees in twitter: data analysis and branching process models. *Social Network Analysis and Mining*, 6(1):1–13, 2016.

[106] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[107] Zizi Papacharissi. The virtual sphere the internet as a public sphere. *New media & society*, 4(1):9–27, 2002.

[108] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[109] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 430–438. ACM, 2011.

[110] Andrew Perrin. Social media usage. *Pew Research Center*, 2015.

[111] Keith T Poole and Howard L Rosenthal. *Ideology and congress*, volume 1. Transaction Publishers, 1997.

[112] Ana-Maria Popescu and Marco Pennacchiotti. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1873–1876. ACM, 2010.

[113] Markus Prior. Media and political polarization. *Annual Review of Political Science*, 16:101–127, 2013.

[114] David P Redlawsk, Andrew JW Civettini, and Karen M Emmerson. The affective tipping point: Do motivated reasoners ever "get it"? *Political Psychology*, 31(4):563–593, 2010.

[115] Sonia Roccas and Marilynn B Brewer. Social identity complexity. *Personality and Social Psychology Review*, 6(2):88–106, 2002.

[116] Daniel M Romero, Brian Uzzi, and Jon Kleinberg. Social networks under stress. In *Proceedings of the 25th International World Wide Web Conference*, pages 9–20, 2016.

[117] Laura M Smith, Linhong Zhu, Kristina Lerman, and Zornitsa Kozareva. The role of social media in the discussion of controversial topics. In *SocialCom*, pages 236–243. IEEE, 2013.

[118] Natalie Jomini Stroud. Media use and political predispositions: Revisiting the concept of selective exposure. *Political Behavior*, 30(3):341–366, 2008.

[119] Natalie Jomini Stroud. *Niche news: The politics of news choice*. Oxford University Press on Demand, 2011.

[120] Cass R Sunstein. The law of group polarization. *Journal of political philosophy*, 10(2):175–195, 2002.

[121] Cass R Sunstein. *Republic. com 2.0*. Princeton University Press, 2009.

[122] Mikalai Tsytsarau, Themis Palpanas, and Kerstin Denecke. Scalable detection of sentiment-based contradictions. *DiversiWeb, International World Wide Web Conference*, 2011, 2011.

[123] Vasileios Tzoumas, Christos Amanatidis, and Evangelos Markakis. A game-theoretic analysis of a competitive diffusion process over social networks. In *WINE*, 2012.

[124] Marshall Van Alstyne and Erik Brynjolfsson. Global village or cyber-balkans? modeling and measuring the integration of electronic communities. *Management Science*, 51(6):851–868, 2005.

[125] VG Vydiswaran, ChengXiang Zhai, Dan Roth, and Peter Pirolli. Overcoming bias to learn about controversial topics. *Journal of the Association for Information Science and Technology*, 2015.

[126] Kevin Wallsten. Political blogs and the bloggers who blog them: Is the political blogosphere and echo chamber. In *American Political Science Association's Annual Meeting.*, pages 1–4, 2005.

[127] Ingmar Weber et al. Political hashtag trends. In *In Proceedings of the 35th European Conference on Information Retrieval*, pages 857–860, 2013.

[128] Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. Quantifying political leaning from tweets, retweets, and retweeters. *TKDE*, 28(8):2158–2172, 2016.

[129] Fang Wu and Bernardo A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, 2007.

Social media and the web have provided a foundation where users can easily access diverse information from around the world. However, over the years, social media has also helped users produce and consume content that only agrees with their beliefs, leading to an increased polarization in the society. The goal of this thesis is to understand the phenomenon of polarization on social media. We develop algorithms to automatically detect polarized topics on social media and propose algorithms that could potentially reduce polarization in the society. Given the ever important role of social media in our lives, as witnessed by recent events such as Brexit and the election of Donald Trump, understanding and countering such phenomenon is the need of the hour.

Aalto-DD 20/2018

ISBN 978-952-60-7832-8 (printed)
ISBN 978-952-60-7833-5 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Sciene
www.aalto.fi