Kiran Garimella Polarization on Social Media

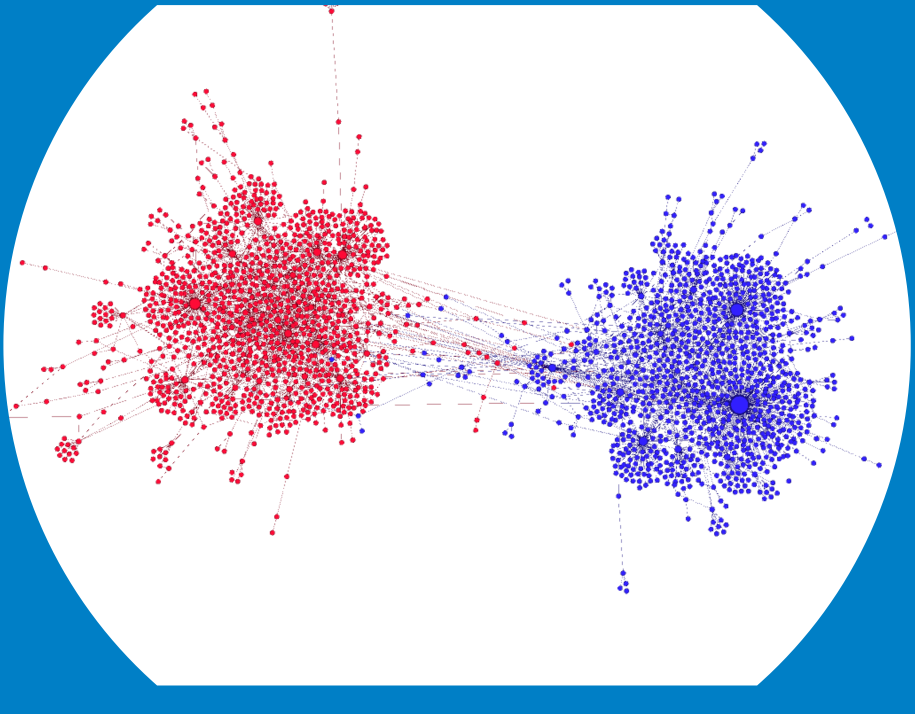
Aalto University

2018

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

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

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.

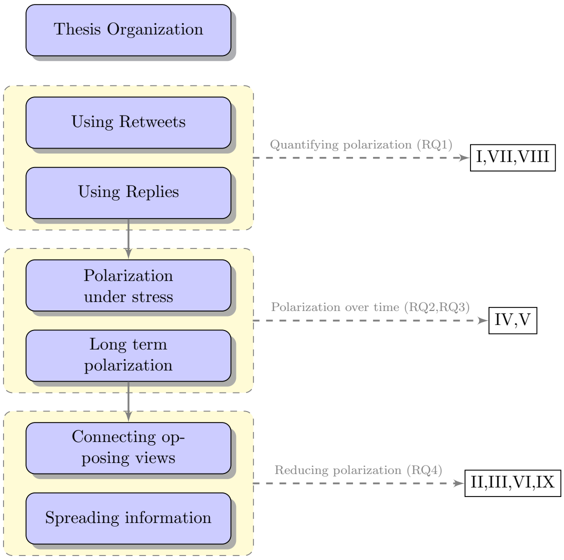

Thesis Organization
Using Retweets
Using Replies
Quantifying polarization (RQ1)
I,VII,VIII
Polarization under stress
Long term polarization
Polarization over time (RQ2,RQ3)
IV,V
Connecting opposing views
Spreading information
Reducing polarization (RQ4)
II,III,VI,IX

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

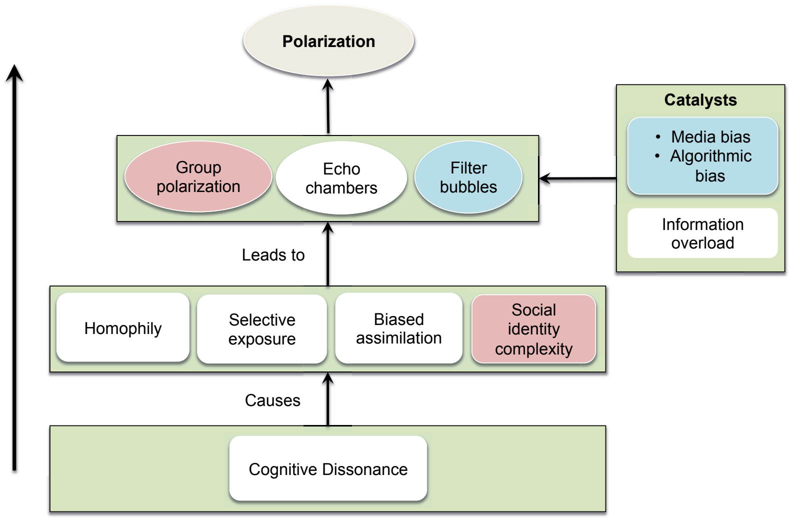


Figure 2.1. Summary of social theories and their dependencies. Individual-level (white), group-level (red) and system-level bias (blue) are colored differently.

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 ($u \rightarrow v$) 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* $\mathscr{T} = (C, R)$, 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 $\mathscr{T}$ 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 $\mathscr{T}$ are enriched with information about publishing time and authoring user. The tree $\mathscr{T}$ can be projected onto the users to model reply interactions among users. The resulting structure is a user *reply graph* $\mathscr{R} = (U, I)$, 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 $u \rightarrow v$ 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*

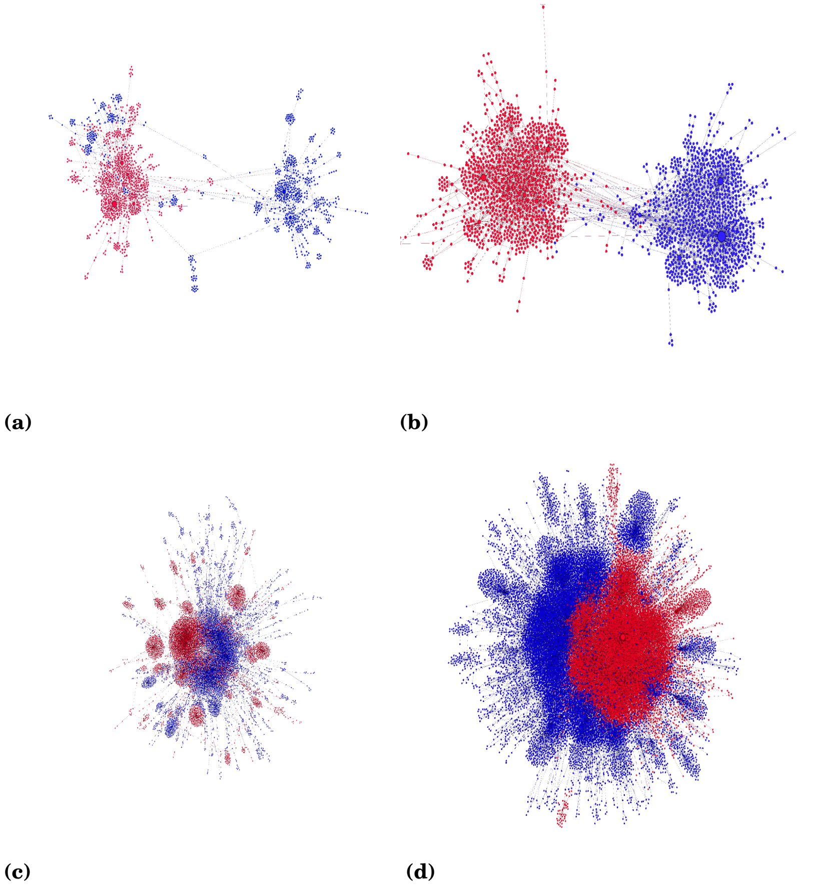


Figure 4.1. Sample retweet graphs (visualized using the force-directed layout algorithm in Gephi). The top two are polarized topics: (a) #beefban, (b) #russia_march, while the bottom are non-controversial, (c) #sxsw, (d) #germanwings. The colors are assigned arbitrarily to the two clusters.

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.

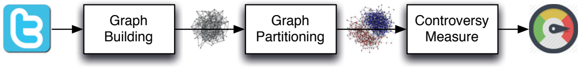


Figure 4.2. Pipeline for quantifying polarization.

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

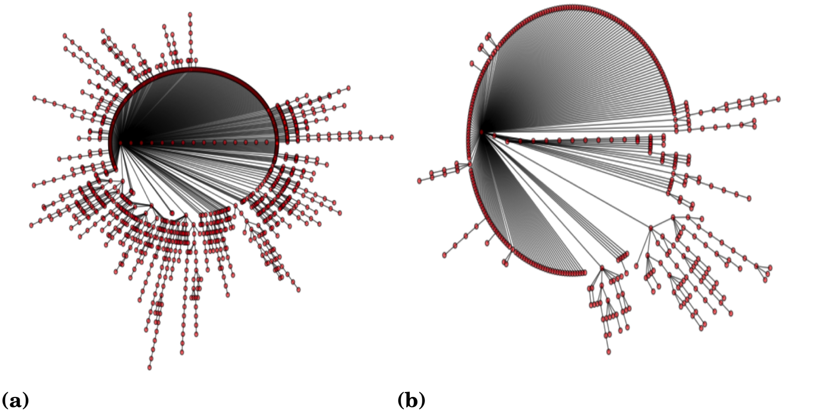


Figure 4.3. Sample reply trees for polarized and non-polarized conversations. All dots start with a root tweet and each subsequent edge in the tree is a reply.

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 $\mathscr{T}$
	num of edges in $\mathscr{T}$
	Avg. degree in $\mathscr{T}$
	Avg. degree in $\mathscr{R}$
propagation	Avg. cascade depth in $\mathscr{T}$
	max cascade depth in $\mathscr{T}$
	max size subtree in $\mathscr{T}$
	Max. relative degree
	Max. degree in $\mathscr{T}$ / root degree in $\mathscr{T}$
	2nd max degree in $\mathscr{R}$
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

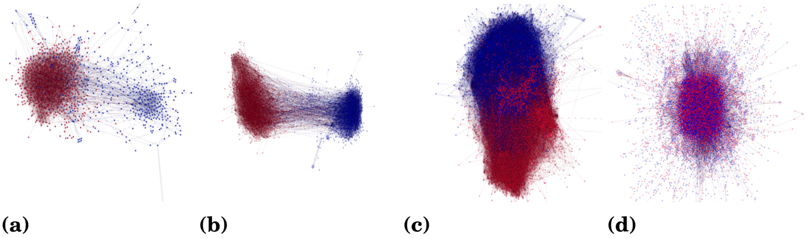


Figure 4.4. Sample follow graphs for polarized topics, (a) #beefban, (b) #russia_march, and non-polarized topics, (c) #sxsw, (d) #germanwings.

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

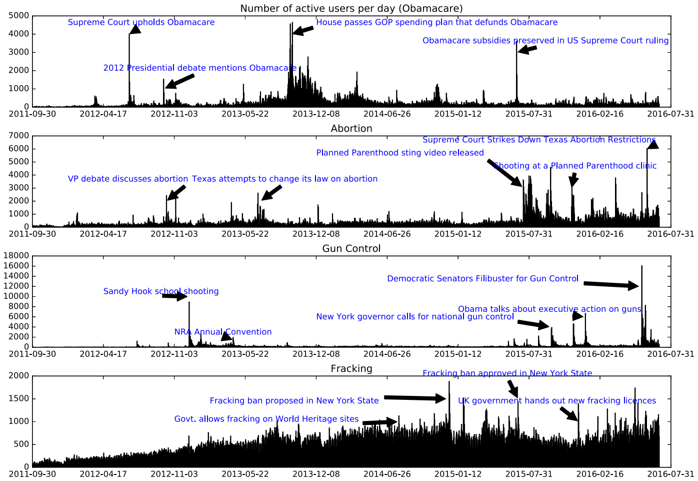


Figure 5.1. Daily trends for number of active users for the four polarized topics under study. Clear spikes occur at several points in the timeline. Manually chosen labels describing related events reported in the news on the same day are shown in blue for some of the spikes.

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.

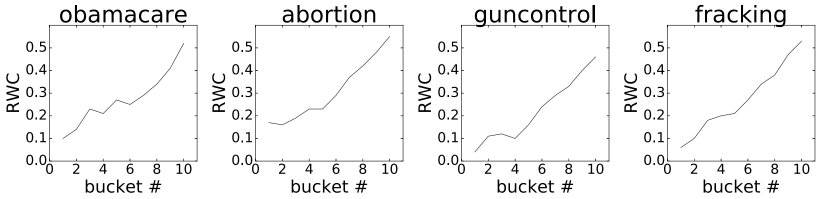


Figure 5.2. RWC score as a function of the activity in the retweet network. An increase in interest in the controversial topic (x-axis) corresponds to an increase in the controversy score of the retweet network.

- 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

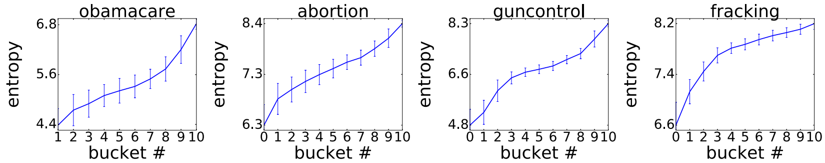


Figure 5.3. Entropy of the distribution over the lexicon for one side of the discussion as a function of the activity in the network (the other side shows similar patterns). As the interest increases, the entropy increases, thus indicating the use of a wider lexicon.

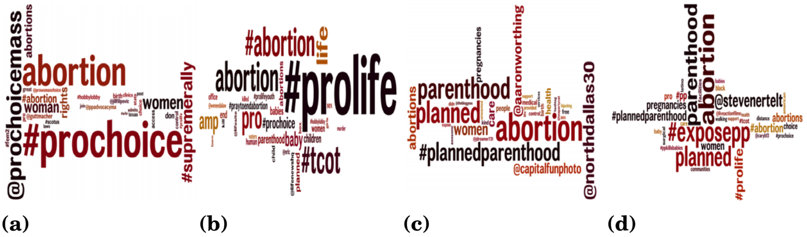


Figure 5.4. Word clouds of content in a discussion on abortion before (a,b) and after (c,d) an event.

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/

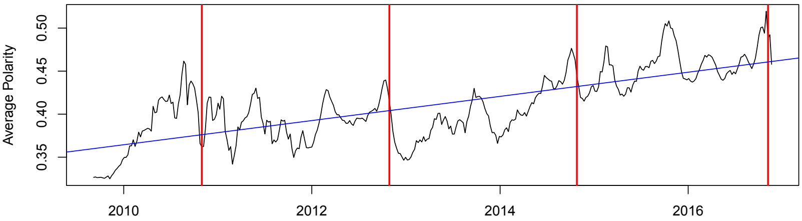


Figure 5.5. Content polarization over time. Red vertical lines indicate the mid term and main elections in the U.S. The blue line is the linear fit, which has a non-zero slope tested using a t-test (p<0.0001).

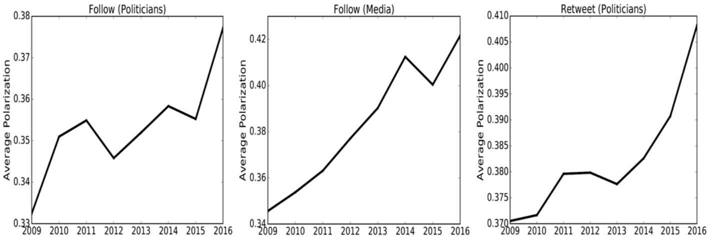


Figure 5.6. Follow (left, middle) and retweet (right) effects over time for politicians (left, right) and media (middle) seed accounts.

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 $X \cap Y = \varnothing$), 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* $\text{RWC}(G', X, Y)$ *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 $X \cap Y = \varnothing$), 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*[$\text{RWC}(G', X, Y)$] *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

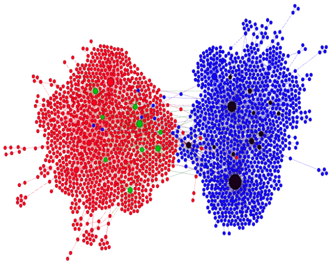


Figure 6.1. An example retweet graph for the topic #russia_march. The green and black dots indicate nodes picked by our algorithm.

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*

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

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.

Publication 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.

Exploring Controversy in Twitter

Kiran
Garimella
Aalto University
Helsinki, Finland
kiran.garimella@aalto.fi

Michael
Mathioudakis
HIIT
Helsinki, Finland
michael.mathioudakis@hiit.fi

Gianmarco
De Francisci Morales
Aalto University
Helsinki, Finland
gdfm@acm.org

Aristides
Gionis
Aalto University
Helsinki, Finland
aristides.gionis@aalto.fi

CSCW '16 Companion, February 27 - March 02, 2016, San Francisco, CA, USA
ACM 978-1-4503-3950-6/16/02.
http://dx.doi.org/10.1145/2818052.2874318

Abstract

Among the topics discussed on social media, some spark more heated debate than others. For example, experience suggests that major political events, such as a vote for healthcare law in the US, would spark more debate between opposing sides than other events, such as a concert of a popular music band. Exploring the topics of discussion on Twitter and understanding which ones are controversial is extremely useful for a variety of purposes, such as for journalists to understand what issues divide the public, or for social scientists to understand how controversy is manifested in social interactions.

The system we present processes the daily trending topics discussed on the platform, and assigns to each topic a *controversy score*, which is computed based on the interactions among Twitter users, and a visualization of these interactions, which provides an intuitive visual cue regarding the controversy of the topic. The system also allows users to explore the messages (tweets) associated with each topic, and sort and explore the topics by different criteria (e.g., by controversy score, time, or related keywords).

Author Keywords

Social media; Controversy

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Introduction

Social media have emerged as the fora of choice for users on the Web to express their opinion about issues they deem important. These fora provide a way to interact with other users who wish to discuss the same issues. Due to their widespread adoption, and the fact that much of the activity they host is publicly available, they offer a unique opportunity to study social phenomena such as peer influence, framing, bias, and controversy. Our work, in particular, is motivated by interest in observing **controversies** at societal level, monitoring their evolution, and possibly understanding which issues become controversial and why.

The system that we demonstrate focuses on the exploration of controversy on Twitter, currently the most popular micro-blogging platform.[1] The back-end of the system processes the messages generated on the platform on a daily basis in order to (i) identify different topics of discussion, (ii) assign a controversy score to each topic, and (iii) produce visual renderings of the activity surrounding each topic in a way that clarifies whether the topic is controversial. The front-end provides a web interface[2] that allows to explore the identified topics according to various views (e.g., ordered by time or magnitude of controversy) and obtain more information about each topic (e.g., by providing a keyword summary of the topic, representative tweets, or a visualization of the activity).

The system is designed to identify controversy on topics in any domain, i.e., without any prior domain-specific knowledge about the topic in question. Specifically, topics are defined by *hashtags*, special keywords conventionally employed by Twitter users to signal that their messages belong to a particular topic. As an example,“#beefban” is a hashtag that was employed to convey that a post referred to a decision of the Indian government, in March 2015, about the consumption of beef meat in India. The system leverages this convention and treats each hashtag as a different topic.

Given a hashtag, we represent the activity on the corresponding topic by a *retweet graph*. In this graph, vertices represent Twitter users who have used the hashtag at least once on a given day, and edges represent *retweets* between users. To quantify the controversy of each topic, we rely on the hypothesis that the structure of the retweet graph reveals how controversial the topic is. This hypothesis is based on the fact that a controversial topic entails different sides with opposing points of view, as well as on previous evidence that individuals on the same side tend to endorse and amplify each other's arguments [1, 2, 3]. We studied this hypothesis in previous work [4], and found strong evidence that the retweet graph of a controversial topic presents a *clustered structure* that reveals the opposing sides of the debate. Moreover, in the same work, we developed a random-walk-based measure that quantifies accurately how controversial a topic is by taking into account the structure of its retweet graph. In light of these findings, for each topic identified, the system computes a controversy score, and produces a rendering of the retweet graph that highlights its clustering structure.

Related Work

Previous studies aim at identifying controversial issues, mostly around political debates [1, 3, 8, 9] but also other topics [5]. While most recent papers focus on Twitter [3, 5, 8, 9], controversy in other social-media platforms, such as blogs [1] and opinion fora [2], has also been analyzed. The main limitation of previous work is that the majority of studies have focused on known, long-lasting debates, such

[1] With 320 million monthly active users as of 30 September 2015 according to https://about.twitter.com/company.

[2] http://users.ics.aalto.fi/kiran/controversy

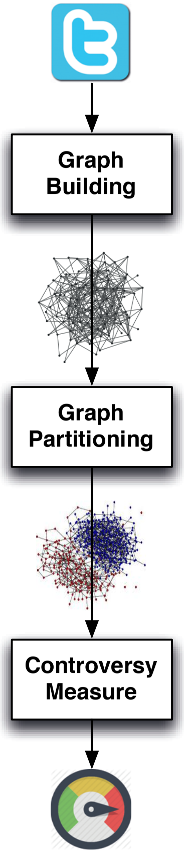


Figure 1: Pipeline for computing controversy scores.

as elections [1, 3]. Our system is the first one to attempt controversy detection in the wild, on any topic, and without human data curation [4].

Quantifying Controversy

Our approach to measuring controversy follows a pipeline with three stages, namely *graph building*, *graph partitioning*, and *measuring controversy*, as depicted in Figure 1. The input to the pipeline is a single hashtag, which defines a topic of discussion. The final output of the pipeline is a value between zero and one that measures how controversial a topic is, with higher values corresponding to higher degree of controversy. We provide a high-level description of each stage here, for further details refer to the original work [4].

Building the Graph

The purpose of this stage is to build the *retweet graph* associated with a single *topic* of discussion. For a given day, each tweet that contains the hashtag that defines the 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, an edge between two vertices signifies that there was one retweet between the corresponding users. We take a retweet as a signal of *endorsement* of opinion between the users.

Partitioning the Graph

In the second stage, the resulting retweet graph is fed into METIS [7] a *graph partitioning* algorithm to extract *two* partitions. 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 controversy– then the two partitions should be only loosely connected to each other, given the semantic of edges.

Measuring Controversy

The third and last stage takes as input the *retweet graph* built by the first stage and partitioned by the second stage, and computes the value of a random-walk-based *controversy measure* [4] that characterizes how controversial the topic is. Intuitively, the controversy measure captures how separated the two partitions are.

Visualizing controversy

As explained in the previous section, we use METIS [7] to produce two partitions on the retweet graph. Given a retweet graph and its two partitions, we produce a visualization of the graph, as in the cases of Figures 2 and . The figures display the two partitions for two topics (#russia_march and #sxsw) in blue and red color on their corresponding retweet graphs. The graph layout is produced by Gephi's ForceAtlas2 algorithm [6], and is based solely on the structure of the graph, not on the partitioning by METIS. It is easy to see that the retweet graph of the first topic is characterized by a bi-modal clustering structure, indicating a controversy. In contrast, the retweet graph of the second topic is characterized by a uni-modal clustering structure, indicating lack of controversy.

Exploration Session

The demonstration will allow the attendees to explore the set of trending topics discussed on Twitter during Jun-Sep 2015. The attendees will be able to interact with a web interface to select one topic and retrieve its *summary*, and organize the set of topics according to different *views*.

Topic Summary. The summary of each topic consists of its hashtag, together with the most related keywords, which

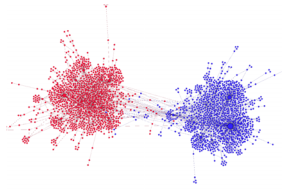


Figure 2: Force directed layout visualization of a controversial topic (#russia_march).

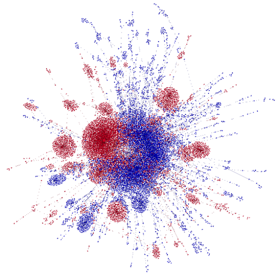
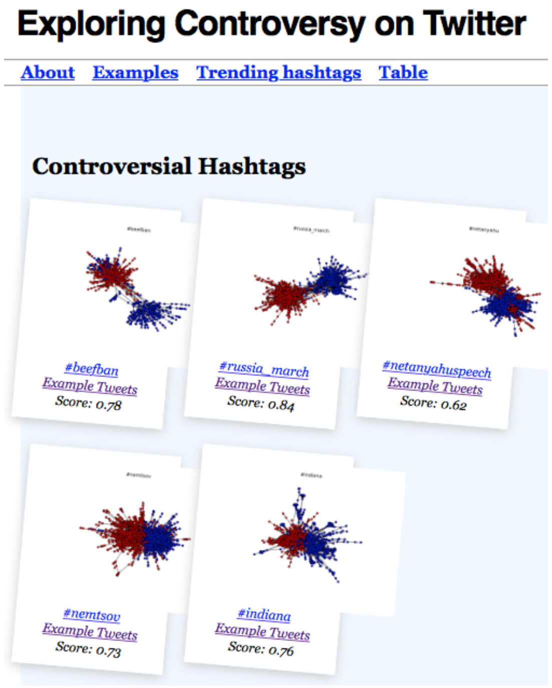


Figure 3: Force directed layout visualization of a non-controversial topic (#sxsw).



Exploring Controversy on Twitter

About Examples Trending hashtags Table

Controversial Hashtags

#beefban
Example Tweets
Score: 0.78

#russia_march
Example Tweets
Score: 0.84

#netanyahuspeech
Example Tweets
Score: 0.62

#nemtsov
Example Tweets
Score: 0.73

#indiana
Example Tweets
Score: 0.76

Figure 4: Screenshot of the web interface.

convey the main idea behind the topic itself. To further help in understanding the topic, the system provides representative tweets from either side of the controversy. These tweets come from authoritative vertices in the graph, as measured by the number of endorsement received. Finally, the summary also includes a visualization of the retweet graph.

Topic Views. Attendees will have the option to browse the topics in chronological order, or sorting them by controversy score, to find the most controversial ones. The system offers also a search functionality, by which the user can specify a text query and obtain a set of relevant topics. Figure 4 is a screenshot of the system while showing some examples of the most controversial topics.

References

[1] Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In *LinkKDD*.

[2] Leman Akoglu. 2014. Quantifying Political Polarity Based on Bipartite Opinion Networks. In *ICWSM*.

[3] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. In *ICWSM*.

[4] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying Controversy in Social Media. In *WSDM*.

[5] Pedro Henrique Calais Guerra, Wagner Meira Jr, Claire Cardie, and Robert Kleinberg. 2013. A Measure of Polarization on Social Media Networks Based on Community Boundaries. In *ICWSM*.

[6] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. (2014).

[7] George Karypis and Vipin Kumar. 1995. Metis - unstructured graph partitioning and sparse matrix ordering system. (1995).

[8] Yelena Mejova, Amy X Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and Sentiment in Online News. *Symposium on Computation + Journalism* (2014).

[9] AJ Morales, J Borondo, JC Losada, and RM Benito. 2015. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos* 25, 3 (2015).

Publication 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.

Reducing Controversy by Connecting Opposing Views

Kiran Garimella
Aalto University
Helsinki, Finland
kiran.garimella@aalto.fi

Gianmarco De Francisci Morales
Qatar Computing Research Institute
Doha, Qatar
gdfm@acm.org

Aristides Gionis
Aalto University
Helsinki, Finland
aristides.gionis@aalto.fi

Michael Mathioudakis
HIIT, Aalto University
Helsinki, Finland
michael.mathioudakis@hiit.fi

ABSTRACT

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 paper we study algorithmic techniques for bridging these chambers, and thus reduce controversy. Specifically, we represent the discussion on a controversial issue with an *endorsement graph*, and 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, which is measured by a recently-developed metric based on random walks. At the same time, we take into account the *acceptance probability* of the recommended edge, which represents how likely the edge is to materialize in the endorsement graph.

We propose a simple model based on a recently-developed user-level controversy score, that is competitive with state-of-the-art link-prediction algorithms. Our goal then becomes finding the edges that produce the largest reduction in the controversy score, in expectation. To solve this problem, we propose an efficient algorithm that considers only a fraction of all the possible combinations of edges. Experimental results show that our algorithm is more efficient than a simple greedy heuristic, while producing comparable score reduction. Finally, a comparison with other state-of-the-art edge-addition algorithms shows that this problem is fundamentally different from what has been studied in the literature.

1. INTRODUCTION

Polarization around controversial issues is a well studied phenomenon in the social sciences [19, 36]. Social media have arguably eased the emergence of such issues, thanks to the scale of discussions and the publicity they foster. This paper studies how to reduce the polarization in controversial issues on social media by creating bridges across opposing sides.

We focus on controversial issues that create discussions online. Usually, these discussions 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*: a vertex v represents a user, and a directed edge (u, v) represents the fact that user u endorses the opinion of user v.

Given this modus operandi, and the existence of confirmation bias, homophily, selective exposure, and related social phenomena in human activities, the existence of echo chambers online is not surprising [13, 8]. The existence of these chambers is a hindrance to the democratic process and to the functioning of society at large, as they cultivate isolation and misunderstanding across sections of the society.

A solution to this problem is to create *bridges* that connect people of opposing views. By putting different parts of the endorsement graph in contact, we hope to reduce the polarization of the discussion the graph represents.

We operationalize this concept by leveraging recent advances in quantifying online controversy [11]. In particular, given a metric that measures how controversial an issue discussed on social media is, our goal is to find a small number of edges, called bridges, that minimize this measure. That is, we seek to propose (content produced by) a user v to another user u, hoping that u endorses v by spreading their opinion. This action would create a new edge (a bridge) in the endorsement graph, thus reducing the controversy score of the graph 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 [22]. 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 a recently developed user-level controversy score [12] which nicely captures the dynamics and properties of the endorsement graph. Therefore, we seek bridges that minimize the *expected* controversy score, according to their acceptance probabilities.

The core of this paper is an algorithm to solve the aforementioned problem. We show that a brute-force approach is not only unfeasible, as it requires one to evaluate a combinatorial number of candidates, but also unnecessary. Moreover, our algorithm needs to consider far fewer than the $\mathcal{O}(n^2)$ possible edges (where n is the number of vertices in the graph) needed by a simple greedy heuristic.

WSDM 2017, February 06 - 10, 2017, Cambridge, United Kingdom

ISBN 978-1-4503-4675-7/17/02. . . $15.00

DOI: http://dx.doi.org/10.1145/3018661.3018703

Experimental results show that our algorithm is able to minimize the controversy score of a graph efficiently and as effectively as the greedy algorithm. In addition, they show that previously-proposed methods for edge addition that optimize for different objective functions are not applicable to the problem at hand.

In summary, our contributions are the following:

- We study the problem of bridging echo chambers algorithmically, in a language and domain agnostic way for the first time. Previous studies that try to address this problem focus mostly on understanding *how* to recommend content to an ideologically opposite side, while our focus is on *who* to recommend contrarian content to. We believe that the two approaches complement each other in bringing us closer to bursting the filter bubble.
- We build on top of results from recent user studies [28, 23, 39] on how users prefer to consume content from opposing views, and formulate the task as an edge-recommendation problem in an endorsement graph, while also taking into account the acceptance probability of a recommendation.
- We provide a method to estimate the acceptance probability of a recommendation that fits well in this setting.
- We propose an efficient algorithm to solve the problem, which considers fewer candidates than a greedy baseline.
- We extensively evaluate the proposed algorithm on real-world data, and demonstrate that it outperforms several sensible baselines.

2. RELATED WORK

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. [21] 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, 7, 25] and quantified [3, 11, 17, 26] this divide.

Given the ill-fated consequences of polarization on society [31, 37], 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 view-points, an idea that has motivated several pieces of work in the literature.

Liao and Fu [22, 23] 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. [39] 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, Munson et al. [28] 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-Garrido et al. [15] 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. [27] 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-Garrido et al. [16] 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, this paper is the first to offer such an algorithmic approach.

(b) Additionally, the studies discussed above focus mostly on understanding *how* to recommend content to an ideologically opposite side. Instead, the approach presented in this paper 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 [28], (*ii*) extreme recommendations might not work [16], (*iii*) people "in the middle" are easier to convince [22], (*iv*) expert users and hubs are often less biased and can play a role in convincing others [23, 39]

In the design of our algorithm we explicitly take into account these considerations (*i*)–(*iv*).

Adding edges to modify the graph structure. In addition to the work on explicitly reducing polarization in social media, there are several papers which aim to make a graph more cohesive by adding edges, where cohesiveness is quantified using graph-theoretic properties such as shortest paths [32, 30], closeness centrality [33], diameter [9], eccentricity [34], communicability [4, 5], synchronizability [40], and natural connectivity [6].

The paper conceptually closest to ours is the one by Tong et al. [38], which aims to add and remove edges in a graph to reduce the dissemination of content (e.g., viruses). The proposed approach maximizes the largest eigenvalue, which determines the epidemic threshold, and thus the properties of information dissemination in networks.

The similarity of the above-mentioned approaches to our paper is limited to the fact that the goal is to modify a graph by edge additions. However, our proposed approach and objective function is predominantly different from those found in other works.

3. PRELIMINARIES AND PROBLEM DEFINITION

To ensure an algorithmic approach to identifying controversial issues and selecting which edges to recommend in order to reduce controversy in a social network, we need to rely on a measure of controversy. As reviewed in Section 2, there are several measures for quantifying controversy in social media [2, 3, 7, 11, 25, 26]. In this paper, we adopt the controversy measure proposed by Garimella et al. [11], as it is the most recent work and it was shown to work reliably in multiple domains; in contrast, other measures focus on a single topic (usually politics) or require domain-specific knowledge. We revise the proposed measure and modify its formulation to adapt it to our current problem. The adopted controversy measure consists of the following steps [11]:

(*i*) Given a topic t for which we want to quantify its controversy level, we create an *endorsement graph* $G = (V, E)$. This graph represents users who have generated content relevant to t. For instance, if t is specified by a hashtag, the vertices of the endorsement graph are the set of all users who have used this hashtag. The edges of the endorsement graph are defined by the *retweets* among the users, in order to capture user-to-user endorsement.

(*ii*) The vertices of the endorsement graph $G = (V, E)$ are partitioned into two disjoint sets X and Y, i.e., $X \cup Y = V$ and $X \cap Y = \varnothing$. The partitioning is based on the graph structure and it is obtained using a graph-partitioning algorithm. The intuition is that, for controversial topics, the partitions X and Y are well separated and correspond to the opposing sides of the controversy.

(*iii*) The last step of computing the controversy measure relies on a *random walk*. In particular, the measure, called *random-walk controversy* (RWC) score, is defined as the difference of the probability that a random walk starting on one side of the partition will stay on the same side and the probability that the random walk will cross to the other side. This measure is computed via two personalized PageRank 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.

In more detail, let P be the column-stochastic transition probability matrix for the random walk, and let X^* and Y^* be the sets of the k_1, k_2 highest in-degree vertices of the two partitions X and Y, respectively. Let r_X be the personalized PageRank vector for the random walk starting in X with restart vector $e_X = \text{uniform}(X)$ and restart probability $(1 - \alpha) \in [0, 1]$, and similarly for r_Y.

Let P_X and P_Y be the transition matrices corresponding to the two random walks starting from the corresponding side. Note that if there are no dangling vertices in the graph then $P_X = P_Y = P$. In the case of dangling vertices, following standard practice, the matrices P_X and P_Y are defined so that the transition probabilities from the dangling vertices are equal to the restart vectors e_X and e_Y, respectively. The personalized PageRank for the two random walks (starting in X and starting in Y) is given by equations:

$$\begin{aligned} r_X &= \alpha P_X r_X + (1 - \alpha) e_X \\ r_Y &= \alpha P_Y r_Y + (1 - \alpha) e_Y . \end{aligned} \tag{1}$$

Let c_X be a vector of size n having value 1 in the coordinates that correspond to the high-degree vertices X^* and 0 elsewhere, and similarly define c_Y. The random-walk controversy score $\text{RWC}(G, X, Y)$ is defined as:

$$\begin{aligned} \text{RWC}(G, X, Y) &= ({c_X}^\intercal r_X + {c_Y}^\intercal r_Y) - ({c_Y}^\intercal r_X + {c_X}^\intercal r_Y) \\ &= (c_X - c_Y)^\intercal (r_X - r_Y). \end{aligned} \tag{2}$$

By using Equations (1), Equation (2) can be written as:

$$\text{RWC}(G, X, Y) = \\ (1 - \alpha)(c_X - c_Y)^\intercal ((I - \alpha P_X)^{-1} e_X - (I - \alpha P_Y)^{-1} e_Y),$$

or

$$\text{RWC}(G, X, Y) = (1 - \alpha)(c_X - c_Y)^\intercal (M_X^{-1} e_X - M_Y^{-1} e_Y), \tag{3}$$

for $M_X = (I - \alpha P_X)$ and $M_Y = (I - \alpha P_Y)$.

Given the controversy measure $\text{RWC}(G)$, the problem we consider in this paper can be formulated 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 $X \cap Y = \varnothing$), 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 $\text{RWC}(G', X, Y)$ is minimized.*

Note that the two partitions X and Y are considered fixed and part of the input. We also consider the high-degree vertices on which the score depends the same in G and G'.

4. ALGORITHMS

A brute-force approach to solve the problem needs to consider all $\mathcal{O}(\binom{n^2}{k})$ combinations of k possible edges to add. A more efficient greedy heuristic would select k edges in k steps, and at each step evaluate the improvement in the value of RWC given by any of the remaining $\mathcal{O}(n^2)$ edges. Even for the greedy approach, though, the number of possible edges to consider is prohibitively large in real settings. Since computation of the controversy score is an expensive operation, we would like to invoke the function as few times as possible. That is, we aim to consider far fewer candidate edges — ideally sub-linear in real-world settings.

At a high level, the algorithm we propose works as follows. It considers only the edges between the high-degree vertices of each side. For each such edge, it computes the reduction in the RWC score obtained when that edge is added to the original graph. It then selects the k edges that lead to the lowest score when added to the graph individually.

Exemplary case

To motivate the proposed algorithm, we study an exemplary case. We use this case to justify our choice to add edges which connect high-degree vertices across the two sides.

Consider a hypothetical directed graph shown in Figure 1. The graph consists of two disjoint stars, each comprised of n vertices. Intuitively, each star represents one side of the controversy. The center of each star is the highest degree vertex of each side. Following the definition of Problem 1 for $k = 1$, we ask which directed edge we should add in order to minimize the controversy score RWC of the entire graph.

Without loss of generality, we consider the following four cases of edges: (i) from a to c, (ii) from a to d, (iii) from b to c, (iv) from b to d. Among these four edges, the first one,

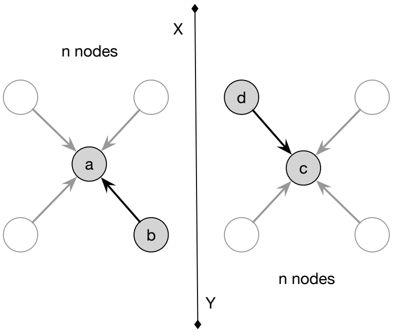


Figure 1: Exemplary case of a graph that consists of two disjoint star-like graphs, each of size n. We wish to add one directed edge so as to minimize the resulting RWC **score.**

from a to c, connects the two centers of the two stars. We can analytically formulate the RWC score we obtain when each of these edges is added to the original graph, denoted with $s_{a\to c}$, $s_{a\to d}$, $s_{b\to c}$, $s_{b\to d}$, respectively. The respective RWC scores are given by the following formulas (details omitted due to lack of space):

$$s_{a\to c} = \frac{(-\alpha^2+\alpha)\cdot n + (\alpha-1)^2}{(\alpha^2+\alpha+1)\cdot n - \alpha^2} + \frac{\alpha\cdot n - \alpha + 1}{(\alpha+1)\cdot n - \alpha}$$

$$s_{a\to d} = \frac{(-\alpha^3+\alpha)\cdot n + \alpha^3 - \alpha^2 - \alpha + 1}{(\alpha^3+\alpha^2+\alpha+1)\cdot n - \alpha^3} + \frac{\alpha\cdot n - \alpha + 1}{(\alpha+1)\cdot n - \alpha}$$

$$s_{b\to c} = \frac{2\alpha\cdot n - 3\alpha + 2}{(\alpha+1)\cdot n - \alpha}$$

$$s_{b\to d} = \frac{\alpha\cdot n - \alpha + 1}{(\alpha+1)\cdot n - \alpha} + \frac{2\alpha\cdot n - 3\alpha - \alpha^2 + 2}{2(\alpha+1)\cdot n + \alpha^2 - 2\alpha}$$

Theorem 1. *For $n \to \infty$, $\alpha \in [0, 1]$, we have*

$$s_{a\to c} \le s_{a\to d}, s_{b\to c}, s_{b\to d}.$$

Proof. We have

$$s_{a\to c} \underset{n\to\infty}{\to} \frac{-\alpha^2+\alpha}{\alpha^2+\alpha+1} + \frac{\alpha}{\alpha+1}$$

$$s_{a\to d} \underset{n\to\infty}{\to} \frac{-\alpha^3+\alpha}{\alpha^3+\alpha^2+\alpha+1} + \frac{\alpha}{\alpha+1}$$

$$s_{b\to c} \underset{n\to\infty}{\to} \frac{2\alpha}{\alpha+1}$$

$$s_{b\to d} \underset{n\to\infty}{\to} \frac{\alpha}{\alpha+1} + \frac{2\alpha}{2(\alpha+1)} = \frac{2\alpha}{\alpha+1}$$

and the inequalitites follow trivially. □

Therefore, the edge from vertex a to vertex c is the one that leads to the minimum score. Theorem 1 provides the optimal edge for a special case. Even though real graphs do not match this case exactly, they often have a structure that resembles star-graphs in certain ways: a small number of highly popular vertices receive incoming edges (endorsements) from a large number of other vertices. We can think that, in a controversial setting, there are thought leaders and followers. Most activity in the endorsement graph happens when spreading the voice of these leaders across their side. This creates a polarized structure which resembles a union of stars on each side of the controversy. The theorem suggests intuitively that edges between high-degree vertices of either side are good candidates to add to obtain a low RWC score.

The exemplary case described above motivates us to consider edges between high-degree vertices from either side. The algorithm for selecting the edges to be added is shown as Algorithm 1. Its running time is $\mathcal{O}(k_1 \cdot k_2)$, where k_1, k_2 are the number of high-degree vertices chosen in X and Y respectively.

Algorithm 1: Algorithm for k-EdgeAddition

```
Input: Graph G, number of edges to add, $k$; $k_1, k_2$ high
       degree vertices in $X, Y$ respectively
Output: List of $k$ edges that minimize the objective
        function, RWC
Initialize: Out ← empty list ;
for $i$ = 1:$k_1$ do
    vertex $u$ = X[i];
    for $j$ = 1:$k_2$ do
        vertex $v$ = Y[j];
        Compute $\delta$RWC$_{u\to v}$, the decrease in RWC if the
          edge (u, v) is added;
        Append $\delta$RWC$_{u\to v}$ to Out;
        Compute $\delta$RWC$_{v\to u}$, the decrease in RWC if the
          edge (v, u) is added;
        Append $\delta$RWC$_{v\to u}$ to Out;
sorted ← sort(Out) by $\delta$RWC by decreasing order ;
return top k from sorted;
```

4.1 Incorporating Acceptance Probabilities

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 $X \cap Y = \varnothing$), 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 $E_{\mathbb{A}}[\mathrm{RWC}(G', X, Y)]$ is minimized under acceptance model $\mathbb{A}$.*

We build such an acceptance model $\mathbb{A}$ on the feature of *user polarity* proposed by Garimella et al. [12]. Intuitively, this polarity score of a user, which takes values in the interval $[-1, 1]$, captures how much the user belongs to either side of the controversy. High absolute values (close to -1 or 1) indicate that the user clearly belongs to one side of the controversy, while central values (close to 0) indicate that the user is in the middle of the two sides. We employ user polarity as a feature for our acceptance model 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 $+1$.

Technically, the polarity score R_u of user u is defined as follows. Let l_u^X and l_u^Y be the expected time a random walk needs to hit the high degree vertices of side X and Y, respectively, starting from vertex u. Moreover, let $\rho^X(u) \in [0,1]$ and $\rho^Y(u) \in [0,1]$ be the fraction of other vertices u' for which $l_{u'}^X < l_u^X$ and $l_{u'}^Y < l_u^Y$, respectively. The polarity of user u is then defined as

$$R_u = \rho^X(u) - \rho^Y(u) \quad \in [-1, 1]. \tag{4}$$

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

$$N_{endorsed}(R_u, R_v)/N_{exposed}(R_u, R_v)$$

where $N_{exposed}(R_u, R_v)$ and $N_{endorsed}(R_u, R_v)$ 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. $N_{exposed}(R_u, R_v)$ is computed by assuming that if v follows u, v is exposed to all content generated by u. In practice, the polarity scores are bucketed to smooth the probabilities. An experimental evaluation in Section 6.2 shows that polarity scores learned this way predict the existence of an edge across datasets with good accuracy.

For a recommended edge (u,v) from vertex u to vertex v, with acceptance probability $p(u,v)$ and RWC decrease $\delta\mathrm{RWC}_{u\to v}$, the *expected decrease* in RWC when the edge is recommended individually is

$$E(u,v) = p(u,v) \cdot \delta\mathrm{RWC}_{u\to v}.$$

Algorithm 1 can be efficiently extended to target the expected RWC decrease by using Fagin's algorithm [10]. Specifically, we take as input two ranked lists of edges (u,v), one ranked by decreasing $\delta\mathrm{RWC}_{u\to v}$ (as currently produced in the course of Algorithm 1) and another one ranked by decreasing probability of acceptance $p(u,v)$. Fagin's algorithm parses the two lists in parallel to find the edges that optimize the expected decrease $E(u,v)$. We refer the interested reader to the original work for details [10].

5. INCREMENTAL COMPUTATION OF RWC

The RWCscore, as defined in Section 3 can be computed via personalized PageRank, which is usually implemented by power iterations. However, since we are only interested in computing the incremental change in RWC after adding an edge, we propose a new way to efficiently compute it.

Consider the transition probability matrix P. After the addition of one (directed) edge from vertex a to vertex b, only one column of P is affected: the column that corresponds to the origin vertex (a) of the directed edge. Let q be the out degree of a. Specifically, before the addition of the edge, the a^{th} column of the matrix has the following form.

$$P^T = \begin{bmatrix} & & & \cdots & & & \\ & & & \cdots & & & \\ \frac{1}{q} & \frac{1}{q} & \ldots & \frac{1}{q} & 0 & 0 & \ldots & 0 \\ & & & \cdots & & & \\ & & & \cdots & & & \end{bmatrix} \tag{5}$$

After adding the new outgoing edge from a, the transition probability matrix has the following form,

$$P^{'T} = \begin{bmatrix} & & & \cdots & & & \\ & & & \cdots & & & \\ \frac{1}{q+1} & \frac{1}{q+1} & \ldots & \frac{1}{q+1} & \frac{1}{q+1} & 0 & \ldots & 0 \\ & & & \cdots & & & \\ & & & \cdots & & & \end{bmatrix} \tag{6}$$

with an additional $\frac{1}{q+1}$ at the b^{th} index, and all other columns of the matrix are unchanged.

Define $u^\mathsf{T} = [0\ 0\ 0\ 0\ \ldots\ 1\ 0\ 0\ \ldots\ 0]$ (the a^{th} vector of the standard basis of $\mathbb{R}^n$). Similarly, define v^T as a column vector with a 1 at the b^{th} position and 0 elsewhere.

Define z^T: (i) If the outgoing vertex a is not a dangling vertex, as $\frac{1}{q+1}\left[\frac{1}{q}\ \frac{1}{q}\ \frac{1}{q}\ \frac{1}{q}\ \frac{1}{q}\ \ldots\ \frac{1}{q}\ -1\ \ldots\ 0\ 0\ 0\right]$ (i.e., $\frac{1}{q(q+1)}$ at all non zero neighbor indices, and $\frac{-1}{q+1}$ at the index of the incoming vertex), we can also say that z_x/z_y is the column vector in P_x/P_y corresponding to the outgoing vertex, multiplied by $\frac{1}{q+1}$ and a -1 at the index of the incoming vertex; and, (ii) if the outgoing vertex is a dangling vertex, as $e_x - v$ or $e_y - v$, depending on the side.

The updated transition probability matrix P' is given by:

$$P^{'} = P - zu^\mathsf{T}. \tag{7}$$

Let $M_x = I - \alpha P_x$ and $M_x^{'} = I - \alpha P_x^{'}$. Expanding the formula for M_x', we get

$$M_x^{'} = I - \alpha P_x^{'} = I - \alpha P_x + \alpha z_x u^\mathsf{T} = M_x + \alpha z_x u^\mathsf{T}. \tag{8}$$

Similarly for $M_y^{'}$, $M_y^{'} = M_y + \alpha z_y u^\mathsf{T}$. As we can see, for any single edge addition, RWCcan be re-computed by using only additional vectors that depends on the vertex that is affected. Moreover, the inverse of $M_x^{'}$ (needed in Equation (3)) can be computed efficiently by using the Sherman-Morrison formula.

Lemma 1 (Sherman-Morrison Formula [14]). *Let* $\mathbf{M}$ *be a square* $n\times n$ *invertible matrix and* $\mathbf{M}^{-1}$ *its inverse. Moreover, let* $\mathbf{a}$ *and* $\mathbf{b}$ *be any two column vectors of size* n*. Then, the following equation holds*

$$(\mathbf{M} + \mathbf{a}\mathbf{b}^T)^{-1} = \mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{a}\mathbf{b}^T\mathbf{M}^{-1}/(1 + \mathbf{b}^T\mathbf{M}^{-1}\mathbf{a}).$$

Now, from Equation (3), the updated RWC, $\mathrm{RWC}^{'}$ is, $\mathrm{RWC}^{'} = (1-\alpha)(c_x - c_y)^\mathsf{T}(M_x^{'-1}e_x - M_y^{'-1}e_y)$, and the update in RWC can be written as

$$\begin{aligned} \delta(\mathrm{RWC}) = \mathrm{RWC}^{'} - \mathrm{RWC} \\ = (1-\alpha)(c_x - c_y)^\mathsf{T}\big((M_x^{'-1}e_x - M_x^{-1}e_x) \\ +(M_y^{-1}e_y - M_y^{'-1}e_y)\big) \\ = (1-\alpha)(c_x - c_y)^\mathsf{T}\Bigg(-\Big(\frac{\alpha M_x^{-1}z_x u^\mathsf{T} M_x^{-1}}{1 + \alpha u^\mathsf{T} M_x^{-1} z_x}\Big)e_x \\ +\Big(\frac{\alpha M_y^{-1}z_y u^\mathsf{T} M_y^{-1}}{1 + \alpha u^\mathsf{T} M_y^{-1} z_y}\Big)e_y\Bigg). \end{aligned} \tag{9}$$

Table 1: Datasets statistics: hashtag used to collect dataset, number of tweets, size of retweet graph.

Dataset	# Tweets	Retweet graph	
		$\lvert V\rvert$	$\lvert E\rvert$
#beefban	84 543	1610	1978
#nemtsov	183 477	6546	10 172
#netanyahuspeech	254 623	9434	14 476
#russia_march	118 629	2134	2951
#indiasdaughter	167 704	3659	4323
#baltimoreriots	218 157	3902	4505
#indiana	116 379	2467	3143
#ukraine	287 438	5495	9452
obamacare	123 320	3132	3241
guncontrol	117 679	2633	2672

In light of Equation (9), the costly inverse computation need not be performed in each iteration to compute the updated RWC score. When a new edge is added to the graph, we just compute the vectors z_x, z_y, and u, and use Equation (9) to directly compute the incremental change in RWC, instead of computing the new RWC and taking the difference. The matrix multiplication $M_*^{-1} z_* u^\intercal M_*^{-1}$ can be computed efficiently by grouping the matrices as $(M_*^{-1} z_*)(u^\intercal M_*^{-1})$. As we see in Section 6.6, this approach provides an order of magnitude speed up in the runtime of our algorithm.

6. EXPERIMENTS

In this section, we provide an evaluation of the two algorithms proposed in Section 4. We use the acronym ROV (*recommend opposing view*) to refer to Algorithm 1, and ROV-AP (*recommend opposing view - with acceptance probability*) to refer to its variation that also considers edge acceptance probabilities.

6.1 Datasets

We use Twitter datasets on known controversial issues. The datasets have also been used in previous studies [11, 24]. Dataset statistics are shown in Table 1. Eight of the datasets consist of tweets collected by tracking single hashtags over a small period of time. The remaining two datasets (*obamacare*, *guncontrol*) consist of tweets collected via the Twitter streaming API[1] by tracking the corresponding keywords for two years. We process the datasets and construct *retweet graphs*. We remark that even though all our datasets are from Twitter, our work can be applied on any graph with a clustered structure separating the sides of a controversy.

6.2 Comparison with other link prediction and recommendation systems

Let us first evaluate the choice of using vertex polarity scores to predict edge acceptance (Section 4.1). To perform this evaluation we compare our approach to other state-of-the-art link-prediction algorithms, which are listed in Table 2.

Following Section 4.1, to estimate acceptance probabilities as a function of user polarity, we first bucket the user polarity scores into 10 equally sized buckets, from -1 to +1. Then, we estimate acceptance probabilities $p(u, v)$ separately for each bucket combination of users u and v. We train a model and cross-validate across all datasets. The median AUC is 0.79, which indicates that endorsement graphs across different datasets have similar edge-formation criteria.

Table 2: Algorithms explored for link prediction.

Algorithm	Summary	AUC
Vertex polarity	Link recommendation based on vertex polarity	0.79
Adamic-Adar [1]	Link prediction based on number of common neighbors	0.60
Reliability [41]	Block stochastic model	0.66
RAI [35]	Using community detection to improve link prediction	0.60
SLIM [29]	Collaborative filtering recommendation	0.71
FISM [20]	Content-based recommendation	0.66

We compare our approach with existing link-recommendation methods. The implementations are obtained from Librec [18]. Table 2 reports the results. As we can see, our approach, which uses vertex polarity scores for predicting links, works as well as the best link-recommendation algorithm. Note that the objective here is not to propose yet another link-recommendation algorithm, nor to claim that our method works better than other approaches in general. Rather, our objective is to validate the use of vertex polarities to create a model for edge-acceptance probabilities.

6.3 Comparison with other related approaches

As mentioned earlier, this paper is the first tp addresses the problem of adding edges to reduce controversy. However, there exist other methods that consider adding edges to improve other structural graph properties. In this section, we compare our approach with three such recent methods: (*i*) NetGel [38], which maximizes the largest eigenvalue; (*ii*) MioBi [6], which maximizes the average eigenvalue; and (*iii*) Shortcut [32], which minimizes the average shortest path. We also experiment with the simple greedy version of our approach, which does not use the heuristic proposed in Section 4, but considers all possible edges.

The results are shown in Figure 2. As expected, the greedy brute-force algorithm performs the best. Our algorithm, ROV, which considers only a small fraction of possible edges, performs quite well, and in some cases, is on par with the greedy. The version of our algorithm with edge acceptance probabilities, ROV-AP, comes next. It is worth noting that even though the choice of edges for ROV-AP is based on a different criterion, the performance of the algorithm in terms of the RWC score is not impacted much. On the other hand, as we will see in Section 6.5, using edge acceptance probabilities improves significantly the real world applicability of our approach.

The other methods (NetGet, MioBi and Shortcut) do not perform particularly well. This is expected, as those methods are not designed to optimize our objective function. Overall, our results demonstrate the need for a specialized method to reduce controversy.

6.4 Edge-addition strategies

Let us now evaluate different edge-addition strategies. The goal is to test the hypothesis that adding edges among high-degree vertices on the two sides of the controversy gives the highest decrease in polarity score. For each of the 10 datasets, we generate a list of random high-degree vertices and non-high-degree vertices on each side. We then generate a list of

[1] https://dev.twitter.com/streaming/public

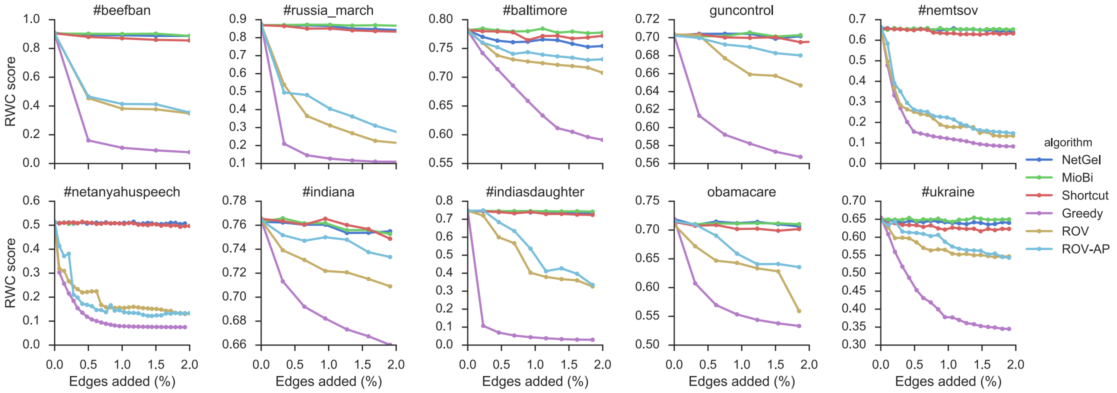


Figure 2: Comparison of the proposed methods (ROV and ROV-AP) with related approaches (NetGel, MioBi, Shortcut) for 2% of the total edges added. The Greedy algorithm considers all possible edges.

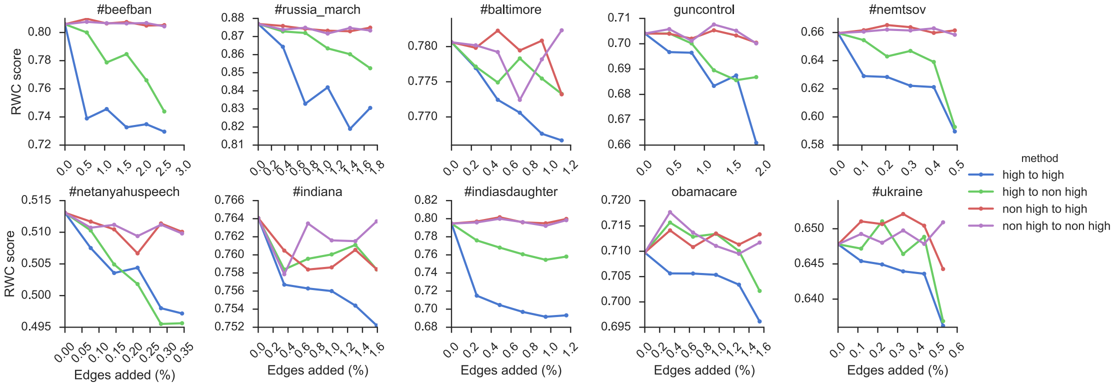


Figure 3: Comparison of different edge-addition strategies after the addition of 50 edges.

50 edges, drawn at random from the sampled vertices, and corresponding to the 4 possible combinations (high/non-high to high/non-high edges). Figure 3 shows the results of these simulations. We see that, despite the fact that high-degree vertices are selected at random, connecting such vertices gives the highest decrease in polarity score (blue line).

6.5 Case study

In order to provide qualitative evidence on the functioning of our algorithms on real-world datasets, we conduct a case study on three datasets. The datasets are chosen for the ease of the interpretation of the results, since they represent topics of wider interest (compared to *beefban*, for example, which is specific to India).

The results of the case study are summarized in Table 3. We can verify that the recommendations we obtain are meaningful and agree with our intuition for the proposed methods. The most important observation is that when comparing ROV and ROV-AP we see a clear difference in the type of edges recommended. For example, for *obamacare*, ROV recommends edges from *mittromney* to *barackbobama*, and from *barackobama* to *paulryanvp* (2012 republican vice president nominee). Even though these edges indeed connect opposing sides, they might be hard to materialize in the real world. This issue is mitigated by ROV-AP, which recommends edges between less popular users, yet connects opposing viewpoints. Examples include the edge (*csgv*, *dloesch*) for guncontrol, which connects a pro-gun-control organization to a conservative radio host, or the edge (*farhankvirk*, *pamelageller*), which connects an islamist blogger with a user who wants to “Stop the Islamization of America.”[2]

Additionally, we provide a quantitative comparison of the output of the two algorithms, ROV and ROV-AP, by extracting several statistics regarding the recommended edges. In particular we consider: (*i*) *Total number of followers*. We compute the median number of followers from all edges suggested by ROV and ROV-AP. A high value indicates that the users are more central. (*ii*) *Overlap of tweet content*, For each edge we compute the Jaccard similarity of the text of the tweets of the two users. We aggregate these values for each dataset, by taking the median among all edges. A higher

[2] Note that since some of the data is from 2012-13, some accounts may have been deleted/moved (e.g., *paulryanvp*, *truthteam2012*). Also, some accounts may have changed stance in these years. Interested readers can use the Internet Archive Wayback Machine to have a look at past profiles.

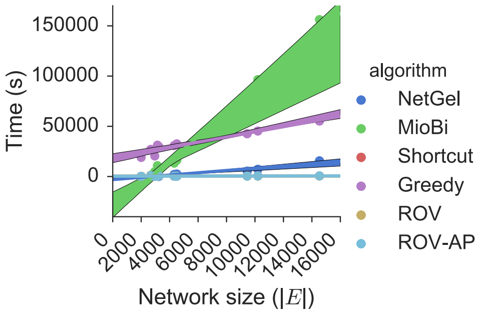


Figure 4: Running time of the proposed algorithms and competitors. ROV **and** ROV-AP **almost overlap.**

value indicates that there is higher similarity between the tweet texts of the two users recommended by the algorithm. (*iii*) *Fraction of common retweets.* For each recommended edge (x, y), we obtain all other users who retweeted users x and y, and compute the Jaccard similarity of the two sets. As before, we aggregate for each dataset, by taking the median among all edges. A higher value indicates that there is a higher agreement in endorsement for users x, y on the topic.

The results are presented in Table 4. We observe that the results agree with our intuition. For example, ROV-AP produces edges with a lower number of followers (not extremely popular users), who have more common retweets, and a higher overlap in terms of tweet content.

6.6 Running time

Finally, we measure the performance of our algorithms in terms of running time. Figure 4 shows that both our algorithms ROV and ROV-AP are fast in comparison to other approaches. Greedy and MioBi are the slowest overall.

Moreover, Figure 5 shows the improvement in running time due to the incremental computation of Section 5. We observe that there is almost an order of magnitude improvement for all the datasets (from 2x – 60x). The density of the graph is indicated by the density of the grey lines in the plot. In general, the speedup is larger for denser graphs.

7. CONCLUSIONS

We considered the problem of bridging opposing views on social media by recommending relevant content to certain users (edges in the endorsement graph). Our work builds on recent studies of controversy in social media and uses a random walk-based score as a measure of controversy. We first proposed a simple, yet efficient, algorithm to bridge opposing sides. Furthermore, inspired by recent user studies on how users prefer to consume content from opposing views, we improved the algorithm to take into account the probability of a recommendation being accepted. Finally, we also proposed a way to incrementally compute the random-walk score by using matrix operations, which typically gives more than an order of magnitude improvement in runtime. We evaluated our algorithms on a wide range of real-world datasets in Twitter, and showed that our methods outperform other baselines.

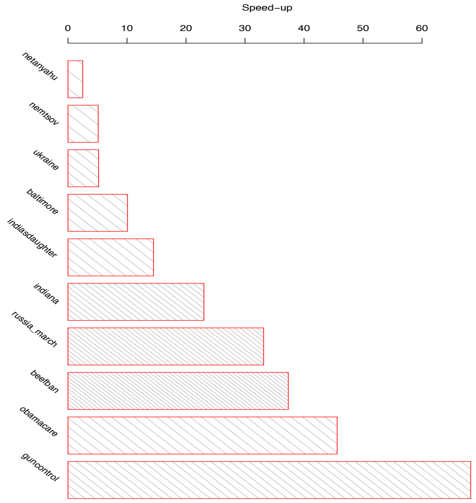


Figure 5: Relative Speed-up produced by using our proposed method in Section 5

Future work. Our approach relies on a random walk-based optimization function [11]. Although this measure has been proven to be effective it has a few drawbacks. In particular, the measure is applicable to controversies having two sides. One way to overcome this restriction is to assume the presence of multiple clusters, and define the measure accordingly. In the future, we plan to experiment with this generalization of our method, as well as, investigate the edge-recommendation problem for other objective functions.

As mentioned in Section 2, previous work deals mostly with the problem of *how* to connect opposing sides, while our work provides methods for selecting *who* to recommend. Another interesting direction is studying the problem of *what* to recommend.

Acknowledgements. This work has been supported by the Academy of Finland project “Nestor” (286211) and the EC H2020 RIA project “SoBigData” (654024).

8. REFERENCES

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

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

[3] L. Akoglu. Quantifying political polarity based on bipartite opinion networks. In *ICWSM*, 2014.

[4] F. Arrigo and M. Benzi. Edge modification criteria for enhancing the communicability of digraphs. *arXiv preprint arXiv:1508.01056*, 2015.

[5] F. Arrigo and M. Benzi. Updating and downdating techniques for optimizing network communicability. *SIAM Journal on Scientific Computing*, 38(1):B25–B49, 2016.

[6] H. Chan, L. Akoglu, and H. Tong. Make it or break it: Manipulating robustness in large networks. In *SDM*, pages 325–333. SIAM, 2014.

Table 3: Twitter handles of the top edges picked by our algorithms for different datasets.

	obamacare		guncontrol		#netanyahuspeech	
	vertex1	vertex2	vertex1	vertex2	vertex1	vertex2
ROV	mittromney realdonaldtrump barackobama barackobama michelebachmann	barackobama truthteam2012 drudge_report paulryanvp barackobama	ghostpanther mmflint miafarrow realalexjones goldiehawn	barackobama robdelaney chuckwoolery barackobama jedediahbila	maxblumenthal bipartisanism harryslaststand lindasuhler thebaxterbean	netanyahu lindasuhler rednationrising marwanbishara worldnetdaily
ROV-AP	kksheld lolgop irritatedwoman hcan klsouth	ezraklein romneyresponse motherjones romneyresponse dennisdmz	chuckwoolery liamkfisher csgv jonlovett drmartyfox	csgv miafarrow dloesch spreadbutter huffpostpol	farhankvirk medeabenjamin 2afight rednationrising jvplive	pamelageller annebayefsky sttbs73 palsjustice chucknellis

Table 4: Quantitative comparison of recommendations from ROV and ROV-AP. * indicates that the result is statistically significant with $p < 0.1$, and ** with $p < 0.001$. Significance is tested using Welch's t-test for inequality of means.

	ROV	ROV-AP
NumFollowers	50729	36160*
ContentOverlap	0.054	0.073**
CommonRetweets	0.029	0.063**

[7] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political Polarization on Twitter. In *ICWSM*, 2011.

[8] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. Echo chambers in the age of misinformation. *arXiv preprint arXiv:1509.00189*, 2015.

[9] E. D. Demaine and M. Zadimoghaddam. Minimizing the diameter of a network using shortcut edges. In *Algorithm Theory-SWAT 2010*, pages 420–431. Springer, 2010.

[10] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. *Journal of computer and system sciences*, 66(4):614–656, 2003.

[11] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy in social media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 33–42, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3716-8. . URL http://doi.acm.org/10.1145/2835776.2835792.

[12] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy in social media. *arXiv preprint arXiv:1507.05224*, 2016.

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

[14] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[15] E. Graells-Garrido, M. Lalmas, and D. Quercia. Data portraits: Connecting people of opposing views. *arXiv preprint arXiv:1311.4658*, 2013.

[16] E. Graells-Garrido, M. Lalmas, and D. Quercia. People of opposing views can share common interests. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 281–282. International World Wide Web Conferences Steering Committee, 2014.

[17] P. H. C. Guerra, W. Meira Jr, C. Cardie, and R. Kleinberg. A measure of polarization on social media networks based on community boundaries. In *ICWSM*, 2013.

[18] G. Guo, J. Zhang, Z. Sun, and N. Yorke-Smith. Librec: A java library for recommender systems. In *Proceedings of the 23rd International Conference on User Modeling, Adaptation and Personalization*, 2015.

[19] D. J. Isenberg. Group polarization: A critical review and meta-analysis. *Journal of personality and social psychology*, 50(6):1141, 1986.

[20] S. Kabbur, X. Ning, and G. Karypis. Fism: factored item similarity models for top-n recommender systems. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 659–667. ACM, 2013.

[21] Q. V. Liao and W.-T. 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.

[22] Q. V. Liao and W.-T. 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.

[23] Q. V. Liao and W.-T. 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.

[24] H. Lu, J. Caverlee, and W. 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.

[25] Y. Mejova, A. X. Zhang, N. Diakopoulos, and C. Castillo. Controversy and sentiment in online news. *Symposium on Computation + Journalism*, 2014.

[26] A. Morales, J. Borondo, J. Losada, and R. Benito. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos*, 25(3), 2015.

[27] S. A. Munson and P. 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.

[28] S. A. Munson, S. Y. Lee, and P. Resnick. Encouraging reading of diverse political viewpoints with a browser widget. In *ICWSM*, 2013.

[29] X. Ning and G. Karypis. Slim: Sparse linear methods for top-n recommender systems. In *2011 IEEE 11th International Conference on Data Mining*, pages 497–506. IEEE, 2011.

[30] M. Papagelis, F. Bonchi, and A. Gionis. Suggesting ghost edges for a smaller world. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2305–2308. ACM, 2011.

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

[32] N. Parotisidis, E. Pitoura, and P. Tsaparas. Selecting shortcuts for a smaller world. In *SIAM International Conference on Data Mining (SDM)*, pages 28–36. SIAM,

2015.

[33] N. Parotsidis, E. Pitoura, and P. Tsaparas. Centrality-aware link recommendations. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 503–512. ACM, 2016.

[34] S. Perumal, P. Basu, and Z. Guan. Minimizing eccentricity in composite networks via constrained edge additions. In *Military Communications Conference, MILCOM 2013-2013 IEEE*, pages 1894–1899. IEEE, 2013.

[35] S. Soundarajan and J. Hopcroft. Using community information to improve the precision of link prediction methods. In *Proceedings of the 21st International Conference on World Wide Web*, pages 607–608. ACM, 2012.

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

[37] C. R. Sunstein. *Republic. com 2.0.* Princeton University Press, 2009.

[38] H. Tong, B. A. Prakash, T. Eliassi-Rad, M. Faloutsos, and C. Faloutsos. Gelling, and melting, large graphs by edge manipulation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 245–254. ACM, 2012.

[39] V. Vydiswaran, C. Zhai, D. Roth, and P. Pirolli. Overcoming bias to learn about controversial topics. *Journal of the Association for Information Science and Technology*, 2015.

[40] A. Zeng, L. Lü, and T. Zhou. Manipulating directed networks for better synchronization. *New Journal of Physics*, 14(8):083006, 2012.

[41] T. Zhou, L. Lü, and Y.-C. Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71 (4):623–630, 2009.

Publication 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.

Mary, Mary, Quite Contrary: Exposing Twitter Users to Contrarian News

Kiran Garimella
Aalto University
Helsinki, Finland
kiran.garimella@aalto.fi

Gianmarco De Francisci Morales
Qatar Computing Research Institute
Doha, Qatar
gdfm@acm.org

Aristides Gionis
Aalto University
Helsinki, Finland
aristides.gionis@aalto.fi

Michael Mathioudakis
Aalto University & HIIT
Helsinki, Finland
michael.mathioudakis@aalto.fi

ABSTRACT

Polarized topics often spark discussion and debate on social media. Recent studies have shown that polarized debates have a specific clustered structure in the endorsement network, which indicates that users direct their endorsements mostly to ideas they already agree with. Understanding these polarized discussions and exposing social media users to content that broadens their views is of paramount importance.

The contribution of this demonstration is two-fold. (*i*) A tool to visualize retweet networks about controversial issues on Twitter. By using our visualization, users can understand how polarized discussions are shaped on Twitter, and explore the positions of the various actors. (*ii*) A solution to reduce polarization of such discussions. We do so by exposing users to information which presents a contrarian point of view. Users can visually inspect our recommendations and understand why and how these would play out in terms of the retweet network.

Our demo[1] provides one of the first steps in developing automated tools that help users explore, and possibly escape, their echo chambers. The ideas in the demo can also help content providers design tools to broaden their reach to people with different political and ideological backgrounds.

1. INTRODUCTION

Social media provides a platform for public discussions on a wide range of topics. Even though users are potentially given a vast choice on what information they can consume, echo chambers, combined with social network design, often limit users to viewpoints that they agree with. Particularly for controversial topics, discussions on social media tend to become polarized, with users supporting their stance and ignoring the one from the opposing side.

Here we present an interactive demo that (*i*) showcases the phenomenon of polarization for discussions on controversial topics, and (*ii*) provides *contrarian* content recommendations, i.e., content that expresses views from the opposing side of the controversy.[2] The goal of this demo is to allow users to explore polarized discussions on social media and, at the same time, show a way to address the polarization phenomenon.

The experience focuses on controversial topics, and especially on the way information is disseminated in these discussions. As previous studies have confirmed, controversial topics create more polarized discussions, which are characterized by specific types of sharing and information dissemination patterns [5]. We model the discussion as an endorsement graph (e.g., retweets in Twitter). In this graph there is an edge between two users u and v if u retweets v.

These polarized discussions, in which people reinforce their existing beliefs, lead to the creation of echo chambers and filter bubbles. Many studies have warned about the threats that these phenomena pose to an open democratic process, as they cultivate isolation and misunderstanding across sections of the society [17].

Our demonstration shows a possible way to address this problem, by exposing people to contrarian content. Differently from previous attempts, our system is fully automated, and employs an algorithm to recommend a set of contrarian news articles that have been shared by the opposing side. By exposing users to content which supports contrarian beliefs, we hope to encourage people to look and understand the point of view of the other side of the controversy.

Our recommendations take into account several factors. The main one comes from recent research in connecting users with opposing views [7]. This factor quantifies the reduction in polarization that a successful recommendation (a recommendation that is endorsed by the recipient) would generate in the discussion.

Clearly, many users might not be interested in content from the other side. For instance, people in the 'center' might be more eager to explore content from either side, as opposed to people on the extremes. To address this concern, our al-

[1] `https://users.ics.aalto.fi/kiran/reducingControversy/homepage`

[2] `https://en.wikipedia.org/wiki/Mary,_Mary,_Quite_Contrary`

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054737

gorithm takes into account each user's history, by computing how likely they are to endorse our recommendation.

The recommendations also take other factors into account, such as the topic distribution and popularity of the content. These factors allow users to get more diverse and engaging content recommendations.

We use several controversial topics ranging over the last two years across multiple domains for this demo, though in practice any controversial topic can be easily incorporated into our demo.

2. RELATED WORK

Even though the Web was envisioned as a place where open discussions on wide range of topics could be facilitated, many users currently do not make use of such an opportunity. Due to phenomena such as homophily, confirmation bias, and selective exposure, people tend to restrict themselves to viewing and sharing information that conforms with their beliefs. Research shows that this phenomenon exists online [11]. This selective exposure has led to increased fragmentation and polarization online. A wide body of recent studies have studied [1, 3, 12] and quantified [2, 4, 5, 10, 13] this issue.

Research also shows that such a division in sections of the society has important consequences to our democracy [17]. There have been attempts to try to nudge users to explore and understand opposing view points. Here we review the most relevant ones.

Wall Street Journal's *Blue feed-Red feed* [3] raises awareness about the extent to which viewpoints on a matter can differ, by showing side-by-side articles expressing very liberal and very conservative viewpoints; *Politecho*[4] displays how polarizing the content on a user's news feed is when compared to their friends'; *Escape your bubble*[5] is a browser extension to add hand-curated content from the opposite side on Facebook; automated bots have been created to respond to posts containing harassment or fake news,[6] with an attempt to de-polarize the discussion and educate users. Moreover, new social media platforms designed to encourage discussions and debates have been proposed, such as (*i*) the Filterburst project,[7] (*ii*) Rbutr,[8] where users can post rebuttals of other urls, and (*iii*) a Wikipedia for debates.[9]

Our demo differs from existing ones in many ways. First, we provide a unique, interactive visualization of an endorsement networks for controversial topics. Second, we showcase a system to recommend contrarian content to users. Our approach is completely algorithmic, unlike most systems listed above, which involve manual curation.

Research has also been done in trying to connect users with opposing views. Munson et al. [14] created a browser widget that measures the bias of users based on the news articles they read. Their study shows that users are willing to slightly change views once they are shown their biases. Graells-Garrido et al. [8] 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. Graells-Garrido et al. [9] propose to find 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. Park et al. [15] propose methods for presenting multiple aspects of news to reduce bias. Garimella et al. [7] study the problem of reducing the overall polarization of a controversial topic in a network. They try to find the best edges to recommend in an endorsement graph so that the polarization score of the entire network is reduced. In this demo, our focus is on reducing the polarization of an individual user (local objective), instead of the entire network (global objective).

3. PRELIMINARIES

A topic of discussion is identified as the set of tweets that satisfy a text query – e.g., all tweets that contain a specific hashtag. We represent a topic with an *endorsement graph* $G(V, E)$, where vertices V represent users and edges E represent *endorsements*.

It has been shown that an endorsement graph captures well the extent to which a topic is controversial [5]. In particular, the endorsement graph of a controversial topic has a *multimodal clustered structure*, where each cluster of vertices represents one viewpoint on the topic. As we focus on two-sided controversies, we identify the two sides of a controversial topic by employing a *graph-partitioning* algorithm, which partitions the graph into *two* subgraphs (represented by X and Y). In this work, we specifically focus on recommending content in the form of news items, such as articles, blog posts, and opinion pieces. The item pool for the recommendation comprises all the links shared by the active users during the observation window.

User polarization score. We use a recently-proposed methodology to define the polarization score for each user in the graph [6]. The score is based on the expected hitting time 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. 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^+ 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

$$\rho(u) = \rho^X(u) - \rho^Y(u) \quad \in (-1, 1). \tag{1}$$

Following this definition, 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

[3] http://graphics.wsj.com/blue-feed-red-feed/

[4] http://politecho.org/

[5] http://www.escapeyourbubble.com/

[6] http://wpo.st/4kVR2 https://goo.gl/Xl6x9t

[7] http://www.filterburst.com/

[8] http://rbutr.com/

[9] http://www.debatepedia.org/en/index.php/Welcome_to_Debatepedia%21

polarization score $\rho(u) \approx 1 - 0 = 1$. The opposite is true for vertices that are far from X^+ but close to Y^+; leading to a polarization score $\rho(u) \approx -1$.

Item polarization score. Once we have obtained polarization scores for users in the graph, it is straightforward to derive a similar score for content items shared by these users. Specifically, we define the polarization score of an item i as the average of the polarization scores of the set of users who have shared i, denoted by U_i:

$$\rho(i) = \frac{\sum_{u \in U_i} \rho(u)}{|U_i|} \quad \in (-1, 1). \tag{2}$$

Acceptance probability. Not all recommendations are agreeable, especially if they do not conform to the user's beliefs. To reduce these effects, we define an acceptance probability, which quantifies the degree to which a user is likely to endorse the recommended content. We use the item and user polarization scores defined above to estimate the likelihood that a target user u endorses (i.e., retweets) the recommended item i. We build an acceptance model by adapting a similar one based on the feature of user polarization [7]. High absolute values of user polarization (close to -1 or 1) indicate that the user belongs clearly to one side of the controversy, while middle-range values (close to 0) indicate that the user is in-between the two sides. It had been shown that users accept content from different sides with different probabilities, and these probabilities can be inferred from the graph structure [7]. For example, a user with polarization close to -1 is more likely to endorse a user with a negative polarization than a user with polarization $+1$. This intuition directly translates to endorsing items, and therefore can be used for our recommendation problem.

Based on this intuition, we define the acceptance probability $p(u, i)$ of a user u endorsing item i as

$$p(u, i) = N_e(\rho(u), \rho(i))/N_x(\rho(u), \rho(i)), \tag{3}$$

where $N_e(\rho(u), \rho(i))$ and $N_x(\rho(u), \rho(i))$ are the number of times a user with polarity $\rho(u)$ has endorsed and was exposed to (respectively) content of polarity $\rho(i)$. In practice, the polarity scores are bucketed to smooth the probabilities.

4. RECOMMENDATION FACTORS

This section describes the factors used to generate content recommendations for the users. 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. Therefore, we take into account several factors: reduction in polarization score of the target user; exclusivity of the candidate items (polarity of the items); acceptance probability of recommendation based on polarization scores; topic diversity; popularity/quality of the candidate item. Next, we describe these factors in more detail.

Reduction of user-polarization score. The maximum reduction of user-polarization score is achieved by putting the user in contact with an authoritative source from the opposing side. Leveraging this idea, we build a list of items L_1, by considering the items shared by high degree nodes on the opposite side of the target user, and ranking them by the potential decrease in polarization score for a user u.

Exclusivity on either side. We consider items that are almost exclusively shared by one of the sides. Specifically, we denote by n_i^X and n_i^Y the number of users who shared each item i on side X and Y, respectively. For each side, we generate a list L_2 ranked by the ratio of shares ${n_i^X}/{n_i^Y}$ (for side X) and ${n_i^Y}/{n_i^X}$ (for side Y).

Acceptance probability. For a given user u, all items sorted in decreasing order of acceptance probability $p(u, i)$ make up list L_3.

Topic diversity. We want to ensure that the recommendations are topically diverse. Therefore, 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 item. Topics are defined as *named entity*, and we extract them via the tool TagMe.[10] Given a user vector t_u, we compute the cosine similarity with all item vectors t_i, and rank items in increasing order of cosine similarity (list L_4).

Popularity on either side. Finally, we 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. List L_5 ranks items by decreasing popularity score.

Rank Aggregation. Given the 5 ranked lists described above, we use a weighted rank-aggregation scheme to generate the final recommendations. The intuition behind using rank aggregation is that items that are highly ranked in many lists, are also highly ranked in the output list. In particular, we use a weighted rank-aggregation technique proposed by Pihur et al. [16]: the goal is to minimize an objective function

$$\phi(\delta) = \sum_{i=1}^{5} w_i d(\delta, L_i), \tag{4}$$

where δ is the optimal ranked output list, d is any distance function (we use the Spearman footrule distance), and w_i are the importance weights of each list. We can set the weights to generate highly contrarian recommendations (by giving large weights to L_1 and L_2) or recommendations that are likely to be accepted (by giving large weight to L_3).

5. ARCHITECTURE

The demo consists of three major parts. (i) Data collection, (ii) Data processing, creation of graphs, and recommendations (detailed in Secions 3 and 4) and (iii) Visualization. We give a brief overview of each of these below.

Data Collection. We collected data from Twitter for eight controversial topics covering a wide range of domains, including the US election results (USElections), and protests against government actions (#baltimoreriots, #beefban, and #nemtsov). Each of these topics contain a few thousands to tens of thousands of users. Though we limit ourselves to these 8 topics in the demo, in practice, a similar methodology can be applied for any controversial topic.

Data Processing. After identifying a controversial topic on Twitter, we construct a *retweet graph*, identify the two sides of the controversy and obtain polarity scores for all users in this topic (detailed in Section 3). Next, for each user, we generate recommendations that surface content

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

with a contrarian point of view. The recommendations for a user are based on their Twitter activity, and take into account five different factors. The details for extracting the recommendations are provided in Section 4.

Visualization. To visualize the results in the demo we use a javascript library called Sigma.js.[11] The retweet network visualizations are created first in Gephi,[12] using a force directed layout and then exported to Sigma.js via a plugin.[13]

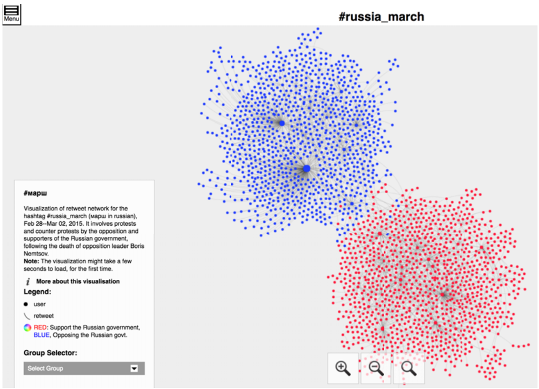
6. DESCRIPTION OF THE DEMO

To use the demo,[14] the viewer first selects a topic from a set of polarized topics on the homepage. Upon selecting a topic, the retweet network corresponding to the topic is shown to the viewer, with the two opposing sides of the controversy highlighted with different colors. On the left there is an info box, where the viewer can find summary information about the topic such as what the discussion is about, why it is polarized, and what the two sides support. The viewer can optionally get more detailed information, including most retweeted tweets on each side, most shared news articles, and other aggregate statistics on the topic, by clicking on the 'More information about this visualization' link. Figure 1 shows the main Web interface for the discussion around the protests for the assassination of Boris Nemtsov in Russia.[15]

The retweet network is at the center of the visualization. The retweet network for controversial topics exhibits a peculiar clustered structure, with two main clusters. By hovering over each node, the viewer can see which other nodes they are connected to. In most cases, nodes are connected to a single side (nodes of the same color), and connections across sides (nodes of different color) are rare. This pattern is an indication that users do not retweet across different sides of the discussion, but only support their own point of view [5]. The viewer can zoom in and out to see specific connections between individual users and groups. Hovering on a node shows their Twitter username, along with their assigned polarity score (higher absolute value means that the user is more polarized). Clicking on a node in the graph highlights the subgraph connected to this node and also brings up an information pane on the right, as shown in Figure 2.

The information pane consists of (*i*) a link to the users profile on Twitter, (*ii*) a sample of three retweets by the user (if the user has retweeted anyone), and (*iii*) three recommendations that aim to expose that specific user to a contrarian viewpoint, along with a set of three random articles. Providing two lists allows the viewer to compare our recommendations to a random baseline recommendation. These samples can be refreshed by clicking on the node again.

For each recommendation, there is a link to show/hide a popup which contains information on *why* that link has been recommended. The popup displays the normalized weights given to the five factors that went into the ranking. Hovering over each recommendation highlights the nodes that have shared this article. This visualization is useful to get an idea on what part of the network shared this article, and hence understand how the recommendation could modify the structure of the endorsement graph of the discussion. Figure 3 shows a depiction of these features.



#марш

Visualization of retweet network for the hashtag #russia_march (марш in russian), Feb 28--Mar 02, 2015. It involves protests and counter protests by the opposition and supporters of the Russian government, following the death of opposition leader Boris Nemtsov.
Note: The visualization might take a few seconds to load, for the first time.

i **More about this visualisation**

Legend:

● user

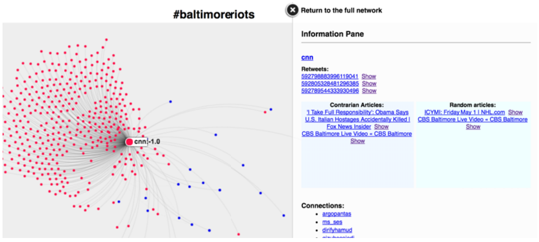
╲ retweet

RED: Support the Russian government, BLUE, Opposing the Russian govt.

Group Selector:

Select Group ▾

Figure 1: Screenshot of the web interface for the topic #russia_march.



Return to the full network

Information Pane

cnn

Retweets:
592798883996119041 Show
592805328481296385 Show
592789544333930496 Show

Contrarian Articles:	**Random articles:**
'I Take Full Responsibility': Obama Says U.S. Italian Hostages Accidentally Killed | Fox News Insider Show CBS Baltimore Live Video « CBS Baltimore Show	ICYMI: Friday May 1 | NHL.com Show CBS Baltimore Live Video « CBS Baltimore Show

Connections:
- argopantas
- ms_ses
- dirifyhamud

Figure 2: Screenshot of the recommendations upon selecting a node (CNN) for #baltimoreriots.

Acknowledgements. This work has been supported by the Academy of Finland project "Nestor" (286211) and the EC H2020 RIA project "SoBigData" (654024).

7. AUTHORS

Kiran Garimella is a PhD student at Aalto University. Previously he worked as a Research Engineer at Yahoo Research, Qatar Computing Research Institute and as an intern at LinkedIn and Amazon. His PhD thesis focuses on identifying, quantifying and combating filter bubbles on social media.

Gianmarco De Francisci Morales is a Scientist at QCRI. Previously he worked as a Visiting Scientist at Aalto University, as a Research Scientist at Yahoo Labs Barcelona, and as a Research Associate at ISTI-CNR in Pisa. He received his Ph.D. in Computer Science and Engineering from the IMT Institute for Advanced Studies of Lucca in 2012. His research focuses on scalable

[11] https://sigmajs.org

[12] https://gephi.org

[13] https://marketplace.gephi.org/plugin/sigmajs-exporter

[14] https://users.ics.aalto.fi/kiran/reducingControversy/homepage/

[15] https://en.wikipedia.org/wiki/Boris_Nemtsov

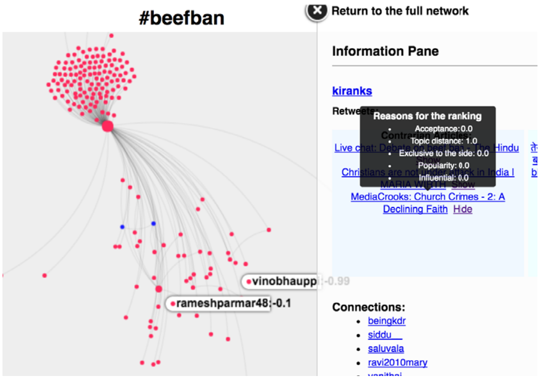


Figure 3: Screenshot showing the weights for a recommendation upon selecting a node, and the users who shared the article on the graph.

data mining, with an emphasis on Web mining and data-intensive scalable computing systems. He is one of the lead developers of Apache SAMOA, an open-source platform for mining big data streams. He co-organizes the workshop series on Social News on the Web (SNOW), co-located with the WWW conference.

Michael Mathioudakis is a Postdoctoral Researcher at Aalto University. He received his PhD from the University of Toronto. His research focuses on the analysis of user generated content on social media, with a recent emphasis on urban computing and online polarization. At Aalto University, he organized and taught new courses on 'Modern Database Systems' and 'Social Web Mining'. He also serves as advisor to Master's students and Aalto's representative at the SoBigData EU project. Outside academia, he worked as a data scientist at Helvia and Sometrik, two data analytics companies.

Aristides Gionis is an associate professor at Aalto University. His research focuses on data mining and algorithmic data analysis. He is particularly interested in algorithms for graphs, social-network analysis, and algorithms for web-scale data. Since 2013 he has been leading the Data Mining Group, in the Department of Computer Science of Aalto University. Before coming to Aalto he was a senior research scientist in Yahoo! Research, and previously an Academy of Finland postdoctoral scientist in the University of Helsinki. He obtained his Ph.D. from Stanford University in 2003.

8. REFERENCES

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

[2] L. Akoglu. Quantifying political polarity based on bipartite opinion networks. In *ICWSM*, 2014.

[3] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political Polarization on Twitter. In *ICWSM*, 2011.

[4] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Exploring Controversy in Twitter. In *CSCW [demo]*, pages 33–36, 2016.

[5] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy in social media. In *WSDM*, pages 33–42. ACM, 2016.

[6] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy in social media. *arXiv preprint arXiv:1507.05224*, 2016.

[7] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Reducing controversy by connecting opposing views. In *WSDM*. ACM, 2017.

[8] E. Graells-Garrido, M. Lalmas, and D. Quercia. Data portraits: Connecting people of opposing views. *arXiv preprint arXiv:1311.4658*, 2013.

[9] E. Graells-Garrido, M. Lalmas, and D. Quercia. People of opposing views can share common interests. In *WWW Companion*, pages 281–282. International World Wide Web Conferences Steering Committee, 2014.

[10] P. H. C. Guerra, W. Meira Jr, C. Cardie, and R. Kleinberg. A measure of polarization on social media networks based on community boundaries. In *ICWSM*, 2013.

[11] Q. V. Liao and W.-T. Fu. Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. In *CHI*, pages 2359–2368. ACM, 2013.

[12] Y. Mejova, A. X. Zhang, N. Diakopoulos, and C. Castillo. Controversy and sentiment in online news. *Symposium on Computation + Journalism*, 2014.

[13] A. Morales, J. Borondo, J. Losada, and R. Benito. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos*, 25(3), 2015.

[14] S. A. Munson, S. Y. Lee, and P. Resnick. Encouraging reading of diverse political viewpoints with a browser widget. In *ICWSM*, 2013.

[15] S. Park, S. Kang, S. Chung, and J. Song. NewsCube: delivering multiple aspects of news to mitigate media bias. In *CHI*, pages 443–452. ACM, 2009.

[16] V. Pihur et al. RankAggreg, an R package for weighted rank aggregation. *BMC bioinformatics*, 10(1):1, 2009.

[17] C. R. Sunstein. *Republic. com 2.0.* Princeton University Press, 2009.

Publication 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.

A Long-Term Analysis of Polarization on Twitter

Kiran Garimella
Aalto University
kiran.garimella@aalto.fi

Ingmar Weber
Qatar Computing Research Institute, HBKU
iweber@hbku.edu.qa

Abstract

Social media has played an important role in shaping political discourse over the last decade. It is often perceived to have increased political polarization, thanks to the scale of discussions and their public nature. Here, we try to answer whether political polarization in the US on Twitter has increased over the last eight years. We analyze a large longitudinal Twitter dataset of 679,000 users and look at signs of polarization in their (i) network - how people follow political and media accounts, (ii) tweeting behavior - whether they retweet content from both sides, and (iii) content - how partisan the hashtags they use are. Our analysis shows that online polarization has indeed increased over the past eight years and that, depending on the measure, the relative change is 10%-20%. Our study is one of very few with such a long-term perspective, encompassing two US presidential elections and two mid-term elections, providing a rare longitudinal analysis.

Introduction

Social media has had a tremendous impact by redefining how we get exposed to information. A recent Pew survey found that more than 60% of Americans get their news from social media (Gottfried and Shearer 2016). While social media has brought about benefits including easier access to knowledge and social connections, social media is also hypothesized to encourage the creation of echo chambers, where users reinforce their own viewpoints and discredit the view points they do not agree with.[1] This can potentially lead to a downward spiral of ever increasing political polarization[2], which, in turn, makes it harder to have a fact-based debate and to reach a consensus on controversial issues.

Though a lot of studies have shown the existence of polarization on social media (Conover et al. 2011; Adamic and Glance 2005), little analysis has been done on long term trends. Performing a study across several decades, as has been done to demonstrate the increasing polarization in the US House of Representatives (Andris and others 2015), is of course impossible as social media is still a fairly recent phenomenon. However, since Twitter was founded in 2006 and its usage now spans several US presidential elections, we still have potential data for a decade.

The main question we want to answer in this paper is if political polarization has increased over time on social media. To address this question, we collect data from Twitter related to both social network structure and tweet content for a large set of users (679,000) engaging with US politics.

We define polarization as a tendency to be restricted in terms of obtaining or engaging with political information to one side of the left-vs.-right political spectrum. To avoid drawing conclusions based on a single perspective, we address three questions each using a different type of information. Namely, (i) have users become less likely to follow both sides of the political spectrum, (ii) have users become less likely to retweet both sides, and (iii) have users become less likely to use hashtags shared by both sides. Our analysis reveals that, according to all three measures, polarization has increased by 10% and 20% between 2009 and 2016.

To the best of our knowledge, this is the first study analyzing political polarization on Twitter over a period of eight years. As it is always easy to get caught up in the heat of the moment, we believe that our study adds a valuable long-term perspective to the evolution of online polarization in the US.

Related Work

Potentially the first study to describe political polarization in a data-driven manner was (Adamic and Glance 2005). The authors point out a clustered structure of hyperlinks between ideologically opposing blogs. Similar analysis on Twitter (Conover et al. 2011) revealed that political polarization exists on Twitter and manifests itself in a way that users endorse (retweet) their own side, but not the opposing side. On a similar note, (Garimella and others 2016) show that most polarized discussions on social media have a well-defined structure, when looking at the retweet network.

From a content perspective, (Mejova and others 2014) consider discussion of controversial and non-controversial news over a span of seven months and identify a correlation between controversial issues and the use of biased and emotional language. They measure bias using manually curated sets of keywords and emotional language using lexicon dictionaries, like SentiWordNet.

In (Weber and others 2013b), the authors study temporal changes in political polarization in Egypt and present

[1] https://goo.gl/SWk1SR

[2] The term *polarization* always refers to *political polarization* in this paper.

evidence that increases in hashtag-based polarization precede events of violence in the real world. In a similar spirit, (Morales et al. 2015) study polarization over a period of two months during the death of Hugo Chavez and identify an increase in polarization in conjunction with external events. Yardi et al. (Yardi and Boyd 2010) study the evolution of a gun violence incident on Twitter for two months and show how homophily plays a role in polarizing discussions. Du et al. (Du and Gregory 2016) compare two separate snapshots of a random sample of the Twitter follower network, one taken in June 2016 and one in August 2016. They then observe that, in line with theories on "triadic closure", 'new edges are (at least 3-4 times) more likely to be created inside existing communities than between communities and existing edges are more likely to be removed if they are between communities'. Such mechanisms could lead to the increase in polarization that we observe in our study.

Perhaps the closest to this work is (Andris and others 2015) who study the partisanship of the US congress over a long period of time. They find that partisanship in the US congress has been increasing for the past few decades.

Concerning longitudinal studies of social media, Liu et al. (Liu and others 2014) analyze seven years of Twitter data to quantify how the users, their behavior, and the site as a whole have evolved. Their work, however, does not describe aspects particular to political polarization.

Dataset

Our dataset is collected around 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 1) for the last eight years. The second type consists of popular media accounts listed in Table 1. The list of media outlets was obtained from a report by the Pew Research Center on polarization and media habits.[3]

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.(Meeder and others 2011). 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 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[4]. 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 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. We perform all our subsequent analysis for retweets and hashtag polarization computation on this data.

Table 1: US 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

Experiments

In our quest to understand long-term polarization trends, we look at three aspects: (i) If user are now more likely to follow users across the political spectrum, (ii) if they are now more likely to retweet such users, and (iii) if users are now more likely to use hashtags which are used by both sides. Here we describe the three types of experiments we performed, related to following and retweeting behavior in the first part, followed by experiments related to hashtags usage.

Following and Retweeting Behavior

To observe changes in the polarization of the following (retweeting) behavior, we wanted to track changes in the probability to follow (retweet) accounts from both sides. As, due to sparsity, following (retweeting) 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 (retweeting) user a uniform prior probability to follow

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

[4] 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

(retweet) seed users from either side. Concretely, we used a beta distribution with a uniform prior (α = β = 1), where α measures the left leaning and β the right leaning.

Then every follow (retweet) 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 (mostly) left leaning user. The mean of the beta distribution, and hence the "leaning" l of the following (retweeting) user, is defined as $l = \alpha/(\alpha + \beta)$, taking the leftness as the direction of the index. 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.

For each political follower/retweeter and each year, this method gives us a value of polarization. Figure 1 plots the distribution of the average polarization and shows temporal shifts. Note that whereas we have the following information for both political and media seed accounts, we only have the retweeting users for political seed. Regardless of whether using politician or media outlet seed accounts, and regardless of whether using following or retweeting information, polarization has increased from 2009 to 2016 between 10-20% in relative terms.

Hashtag Polarization

The third type of polarization we analyze relates to content polarization, more specifically to the polarization of hashtags used by users. Conceptually, a society could be thought of as polarized if there are two opposing sides who speak different languages, in that they differ completely in the words they choose to describe things. For example, one person's "global warming" could be another person's "climategate". We operationalize this idea by applying the methodology previously used in (Weber and others 2013a).

In their methodology, for a given week, a user is assigned a leaning based on the political seed accounts they retweet during that week. Users not retweeting seed accounts during the week do not contribute. Each hashtag h is then assigned a leaning l_h between 0.0=right and 1.0=left based on the leaning of the users using the hashtags in the given week. Differences in user volumes for the two sides are corrected for and smoothing is applied to deal with sparsity. For each hashtag h in a given week, its polarization p_h is then, as before, defined as $p_h = 2 \cdot |0.5 - l|$. The values of p_h are then averaged across all h used in a given week by retweeting users. See (Weber and others 2013a) for details.

To further reduce noise due to low volumes, in particular during the early years, we (i) ignored hashtags used by fewer than five users, and (ii) computed moving averages across five weeks. To look for drifts in the time series, we first tested for stationarity of the time series. An augmented Dickey-Fuller test found the time series to be non-stationary ($p < 0.0001$). Next, we computed the linear fit across time and tested the value of the non-zero slope for statistical significance using a t-test ($p < 0.0001$).

Figure 2 shows the temporal changes for the measure of hashtag polarization together with the linear fit superimposed. Similar to the following and retweeting polarization, there is a relative increase of about 20% between 2009 and 2016. Due to the finer time scale, Figure 2 also suggests that the time around elections corresponds to local maxima in polarization, whereas the time after elections corresponds to local minima. For the 2010 midterm elections, however, this observation does not hold, potentially due to noisier estimates based on fewer active users on Twitter.

Conclusions

There is conflicting evidence on whether social media (i) actively increases offline polarization through the formation of online echo chambers, (ii) merely reflects offline polarization (Vaccari and others 2016), or (iii) helps to *reduce* offline polarization by exposing users to a more diverse set of opinions than they would find in their offline social network (Bakshy and others 2015).

Though our analysis does not directly settle this debate, it provides evidence that polarization on Twitter has increased over the past eight years, potentially reflecting increases in offline polarization as those observed in the US House of Representatives (Andris and others 2015). Furthermore, for three different methodologies, the relative size of the increase of polarization was found to be between 10% and 20%.

In our work, we did not explicitly attempt to detect "astroturfing" and other types of automated tweets (Ratkiewicz and others 2011). However, due to the longitudinal nature of our study, by the time of the data collection (2016/2017) Twitter will have had time to catch most cases of users violating their terms of service, suspending their accounts. More organic efforts such as hashtag hijacking (Hadgu and others 2013) could still affect our analysis, though these effects are arguably also part of the political landscape and should be included.

Given the running up to the 2016 US presidential elections and concerns of a President Trump - famous for his Twitter politics - doing little to attempt to reduce polarization, we speculate that the online polarization will continue to increase in the foreseeable future.

Finally, as social media is coming of age and soon teenagers will no longer remember a pre-social-media era, new opportunities for longitudinal studies arise. At the same time, technical challenges related to "most recent activity only" API limitations hinder such studies. Still, we expect and look forward to more long-term analysis such as ours in the future.

References

Adamic, L. A., and Glance, N. 2005. The political blogosphere and the 2004 us election: divided they blog. In *LinkKDD*, 36–43.

Andris, C., et al. 2015. The rise of partisanship and super-cooperators in the us house of representatives. *PloS one* 10(4):e0123507.

Bakshy, E., et al. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science* 348(6239):1130–1132.

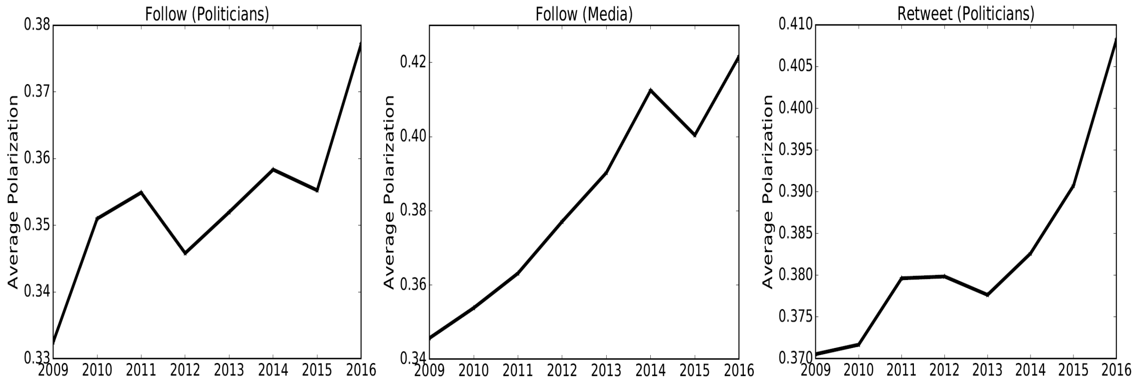


Figure 1: Follow (left, middle) and retweet (right) effects over time for politicians (left, right) and media (middle) seed accounts.

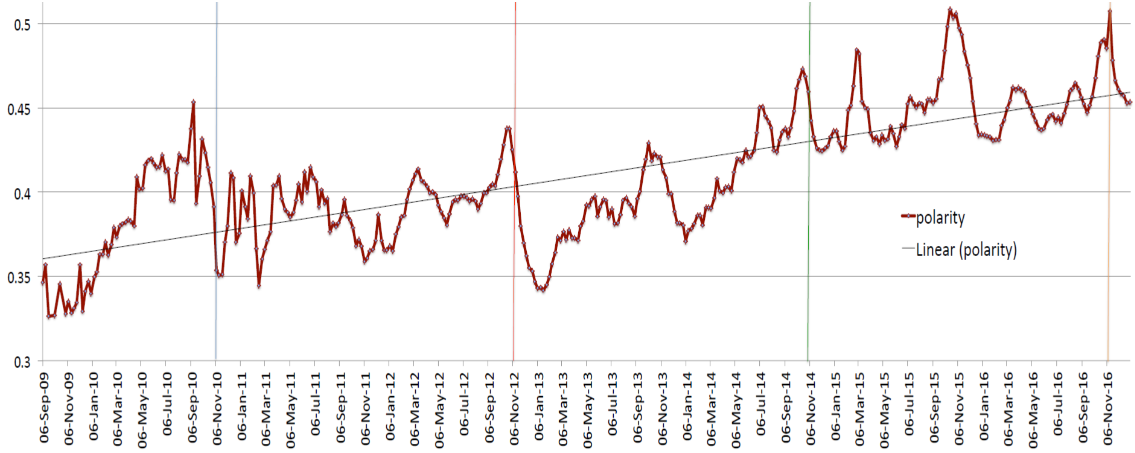


Figure 2: Weekly hashtag polarization. The four vertical lines indicate the 2010 midterm, the 2012 presidential, the 2014 midterm, and the 2016 presidential elections.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011. Political Polarization on Twitter. In *ICWSM*.

Du, S., and Gregory, S. 2016. The echo chamber effect in twitter: does community polarization increase? In *International Workshop on Complex Networks and their Applications*, 373–378.

Garimella, K., et al. 2016. Quantifying Controversy in Social Media. In *WSDM*, 33–42.

Gottfried, J., and Shearer, E. 2016. News use across social media platforms 2016. *Pew Research Center*.

Hadgu, A. T., et al. 2013. Political hashtag hijacking in the us. In *WWW*, 55–56.

Liu, Y., et al. 2014. The tweets they are a-changin: Evolution of twitter users and behavior. In *ICWSM*.

Meeder, B., et al. 2011. We know who you followed last summer: inferring social link creation times in twitter. In *WWW*, 517–526.

Mejova, Y., et al. 2014. Controversy and sentiment in online news. *Symposium on Computation + Journalism*.

Morales, A.; Borondo, J.; Losada, J.; and Benito, R. 2015. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos* 25(3).

Ratkiewicz, J., et al. 2011. Detecting and tracking political abuse in social media. In *ICWSM*, 297–304.

Vaccari, C., et al. 2016. Of echo chambers and contrarian clubs: Exposure to political disagreement among german and italian users of twitter. *Social Media+ Society* 2(3):2056305116664221.

Weber, I., et al. 2013a. Political hashtag trends. In *ECIR*, 857–860.

Weber, I., et al. 2013b. Secular vs. islamist polarization in egypt on twitter. In *ASONAM*, 290–297.

Yardi, S., and Boyd, D. 2010. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society* 30(5):316–327.

Publication 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.

The Effect of Collective Attention on Controversial Debates on Social Media

Kiran Garimella
Aalto University
Helsinki, Finland
kiran.garimella@aalto.fi

Gianmarco De Francisci Morales
Qatar Computing Research Institute
Doha, Qatar
gdfm@acm.org

Aristides Gionis
Aalto University
Helsinki, Finland
aristides.gionis@aalto.fi

Michael Mathioudakis
Aalto University
Helsinki, Finland
michael.mathioudakis@aalto.fi

ABSTRACT

We study the evolution of long-lived controversial 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. 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.

This work is the first to study the dynamic evolution of polarized online debates at such scale. By employing a wide array of network and content analysis measures, we find consistent evidence that increased collective attention is associated with increased network polarization and network concentration within each side of the debate; and overall more uniform lexicon usage across all users.

1 INTRODUCTION

Social media are a major venue of public discourse today, hosting the opinions of hundreds of millions of individuals. Due to their prevalence they have become an invaluable instrument in the study of social phenomena and a fundamental subject of *computational social science*. In this work, we study discussions around issues that are deemed important at a societal level — and in particular, ones that are controversial. This work is a step towards understanding how the discussion about controversial topics on social media evolves, and more broadly how these topics shape the discussion at a societal and political level [1, 16, 23].

We study how online discussions around controversial topics change as interest in them increases and decreases. We are motivated by the observation that interest in enduring controversial issues is re-kindled by external events, e.g., when a major related story is reported. One typical example is the gun control debate in U.S., which is revived whenever a mass shooting occurs.[1] The occurrence of such an event commonly causes an increase in collective attention, e.g., in volume of related activity in social media.

Given a controversial topic, our focus is to analyze the interactions among users involved in the discussion, and quantify how certain structural properties of the interaction network vary with the change in volume of activity. Our main finding is that the polarization reflected in the network structure of online interactions is correlated with the increase in the popularity of a topic.

Differently from previous studies, we study the *dynamic* aspects of controversial topics on social media. While the evolution of networks and polarization on social media have been studied in the past [7, 21], they have not been studied in conjunction before. In addition, we seek to understand the *response* of social media to stimuli that cause increased interest in the topics, an issue that only very recently has seen some attention [30].

We take a longitudinal approach and collect data from Twitter that covers approximately five years. This dataset gives us a very fine-grained view of the activity on social media, including the structure of the interactions among users, and the content they produced during this period. We track four topics of discussion that are controversial in the U.S., that are recurring, and have seen considerable attention during the 2016 U.S. elections.

Our methodology relies on recent advances in quantifying controversy on social media [12]. We build two types of networks: an *endorsement* network from the retweet information on Twitter, and a *communication* network from the replies. We aggregate the data at a daily level, thus giving rise to a time series of interaction graphs. Then, we identify the sides of a controversy via graph clustering, and find the *core* of the network, i.e., the users who are consistently participating to the online discussion about the topic. Finally, we employ a wide array of measures that characterize the discussion about a topic on social media, both from the point of view of the network structure and of the actual content of the posts.

Apart from our main result — an increase in polarization linked to increased interest — we also report on several other findings. We find that most of the interactions during events of interest happen within the different controversy sides, and replies do not cross sides very often, in line with previous observations [31]. In addition, increased interest does not alter the fundamental structure

Contact authors: Kiran Garimella, Michael Mathioudakis.

WebSci'17, June 25-28, 2017, Troy, NY, USA.

978-1-4503-4896-6/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3091478.3091486

[1] See, e.g., http://slate.me/1NswLLD.

of the endorsement network, which is hierarchical, with a disproportionately large fraction of edges linking the periphery to the core. This finding suggests that most casual users, who seldom participate in the discussion, endorse opinions from the core of the side they belong to. When looking at the content of the posts on the two sides of a controversy, we find a consistent trend of *convergence*, as the lexicons become both more uniform and more similar to each other. This result indicates that, while the discussion is still controversial, both sides of the debate focus over the same fundamental issues brought under the spotlight by the event at hand. Conversely, we do not find a consistent long-term trend in the polarization of discussions, which contradicts the common narrative that our society is becoming more divided over time. Finally, we perform similar measurements for a set of topics that are non-political and non-controversial, and highlight differences with the results for controversial discussions.[2]

2 RELATED WORK

A few studies exist on the topic of controversy in online news and social media. In one of the first papers, Adamic and Glance [2] study linking patterns and topic coverage of political bloggers, focusing on blog posts on the U.S. presidential election of 2004. 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 a more recent study of Conover et al. [7], who focus on 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. In another recent work, Mejova et al. [26] consider discussions of controversial and non-controversial news over a span of 7 months. They find a significant correlation between controversial issues and the use of negative affect and biased language. More recently, Garimella et al. [12] show that controversial discussions on social media have a well-defined structure, when looking at the *endorsement* network. They propose a measure based on random walks (RWC), which is able to identify controversial topics, and *quantify* the level of controversy of a given discussion via its network structure alone.

The aforementioned studies focus on static networks, which are a snapshot of the underlying dynamic networks. Instead, we are interested in network dynamics and, specifically, in how it responds to increased collective attention in the controversial topic.

Several studies have looked at how networks evolve, and proposed models of network formation [20, 21]. Densification over time is a pattern often observed [21], 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. [29] offer a comprehensive review. 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 and Mendes [10]. 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. [9] report on a user study that shows how the network structure affects the formation of stereotypes when discussing controversial 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.

Garimella and Weber [15] study polarization on Twitter over a long period of time, using content and network-based measures for polarization and find that over the past decade, polarization has increased. We find no consistent trend among the topics we study.

Perhaps the closest work to this paper is the work by Smith et al. [31], 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.

However, there are differences with our work: (i) they study one local topic (California ballot), over a small period of time, while we study a wide range of popular topics, spanning multiple years; and (ii) their analysis is mostly user centric, whereas we take a global viewpoint, constructing and analyzing networks of user interaction.

The effect of external events on social networks. A few studies have examined the effects of events on social networks. Romero et al. [30] 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 issues. Given the different setting, many of our findings are quite different.

Other works, such as the ones by Lehmann et al. [19] and Wu and Huberman [32], examine how collective attention focuses on individual topics or items and evolves over time. Lehmann et al. [19] 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 and Huberman [32], on the other hand, examine items posted on digg.com and how their popularity decreases over time.

Morales et al. [27] study polarization over time for a single event, the death of Hugo Chavez. Our analysis has a more broad 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.

[2] A limited subset of our results appeared in a poster at ICWSM 2017 [14].

Table 1: Keywords for the controversial topics.

Topic	Keywords	#Tweets	#Users
Obamacare	obamacare, #aca	866 484	148 571
Abortion	abortion, prolife, prochoice, anti-abortion, pro-abortion, planned parenthood	1 571 363	327 702
Gun Control	gun control, gun right, pro gun, anti gun, gun free, gun law, gun safety, gun violence	824 364	224 270
Fracking	fracking, #frack, hydraulic fracturing, shale, horizontal drilling	2 117 945	170 835

Andris et al. [4] 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. Our study suggests that increased controversy is linked to an increase in attention on a topic, whereas we do not see a global trend over time.

3 DATASET

Our study uses data collected from Twitter. Using the repositories of the Internet Archive,[3] we collect a 1% sample of tweets from September 2011 to August 2016,[4] for four topics of discussion, related to 'Obamacare', 'Abortion', 'Gun Control', and 'Fracking'. These topics constitute long-standing controversial issues in the U.S.[5] and have been used in previous work [25]. For each topic, we use a keyword list as proposed by Lu et al. [25] (shown in Table 1), and extract a base set of tweets which contain at least one topic-related keyword. To enrich this original dataset, we use the Twitter REST API to obtain all tweets of users who have participated in the discussion at least once.[6] Admittedly, this dataset might suffer from sampling bias, however the topics are specific enough that the distortion should be negligible [28]. There might also be recency bias due to the addition of the latest tweets of the users. However, the data does not show any clear trend in this sense (see Figure 1). In addition, given that we rely on detecting volume peaks, the trend does not affect our analysis. Table 1 shows the final statistics for the dataset.

We infer two types of interaction network from the dataset: (*i*) a retweet network — a directed endorsement network of users, where there is an edge between two users ($u \rightarrow v$) if u retweets v, and (*ii*) a reply network — a directed communication network of users, where an edge ($u \rightarrow v$) indicates that user u has replied to a tweet by user v. Note that replies are characterized by a tweet starting with '`@username`' and do not include mentions and retweets.[7]

Polarized networks, especially the ones considered here, can be broadly characterized by two opposing *sides*, which express different opinions on the topic at hand. It is commonly understood that retweets indicate endorsement, and endorsement networks for controversial topics have been shown to have a bi-clustered structure [7, 12], i.e., they consist of two well-separated clusters that correspond 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 [6, 24]. These two types of network capture different dynamics of activity, and allow us to tease apart the processes that generate these interactions.

In this paper, we build upon the observation that the clustering structure of retweet networks reveals the opposing sides of a topic. In particular, following an approach from previous work [12], we collapse all retweets contained in the dataset of each topic into a single large static retweet network. Then, we use the METIS clustering algorithm [17] to identify two clusters that correspond to the two opposing sides. This process allows us to identify more consistent sides for the topic. We evaluate the sides by manual inspection of the top retweeted users, URLs, and hashtags. The results are consistent and accurate, and can be inspected online.[8]

Let us now consider the temporal dynamics of these interaction networks. Given the traditional daily news reporting cycle, we build the time series of 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 1. These spikes usually correspond to external newsworthy events, as shown by the annotations. These results support the observation that Twitter is used as an *agorá* to discuss the daily matters of public interest [8].

As shown in Figure 1, the size of the active network for each day varies significantly. There is, however, a *hard core* set of active users who are involved in the discussion of these controversial topics most of the time. 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. Specifically, to build a *core* set of users, we first identify two subsets — one consisting of those users who generated or received a retweet at least once per month for 45 months; and another one defined similarly for replies. We define the core set of users as the union of the aforementioned two sets. Nodes of a network that do not belong to the core are said to belong to the *periphery* of that network. The size of the core ranges from around 600 to 2800 nodes for the four topics. For any given day, the core accounts for at most around 10% of the active users.

3.1 Notation

The set of retweets that occur within a single day d gives rise to one retweet network N_d^{rt}. Each user associated with a retweet is represented with one node in the network. There is a directed edge from user u to user v only when user u has retweeted at least one tweet authored by user v. Correspondingly, the set of replies that occur within a single day give rise to a reply network N_d^{re}. In addition, each node u in the network is associated with a binary

[3] https://archive.org/details/twitterstream

[4] To be precise, we have data for 57 months from that period

[5] According to http://2016election.procon.org.

[6] Up to 3200 due to limits imposed by the Twitter API.

[7] See also https://support.twitter.com/articles/14023 for terminology related to different types of Twitter messages.

[8] https://mmathioudakis.github.io/polarization/

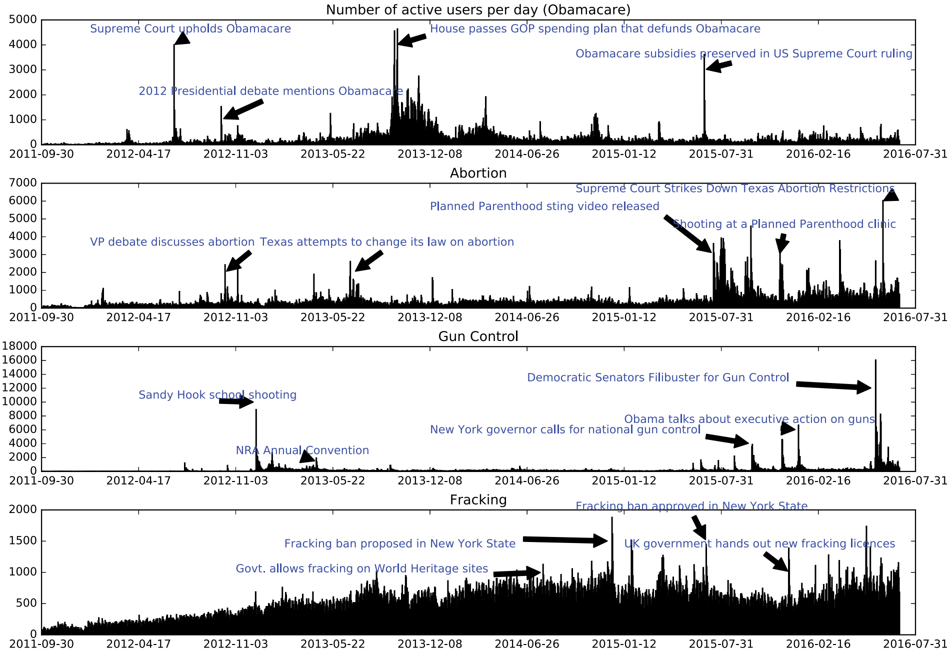


Figure 1: Daily trends for number of active users for the four controversial topics under study. Clear spikes occur at several points in the timeline. Manually chosen labels describing related events reported in the news on the same day are shown in blue for some of the spikes.

attribute $c(u) \in \{\texttt{true}, \texttt{false}\}$ that indicates whether the node is part of the core, and an attribute $s(u) \in \{1, 2\}$ that represents the side the node belongs to. In some cases, we consider undirected versions of the networks defined above. In such cases, we write G_d^{rt}, G_d^{re} to denote the undirected graphs corresponding to N_d^{rt}, N_d^{re}, respectively.

Besides these two types of network, for each day we consider the set of tweets that were generated on that day. Every tweet m is associated with an attribute $s(m) \in \{1, 2\}$ that indicates the side its author belongs to. Moreover, every tweet m is associated with the list of words $w(m)$ that occur in its text. This information gives rise to two unigram distributions W_d^1 and W_d^2, one for each side. Each distribution expresses the number of times each word appears in the tweets of nodes from each side.

4 MEASURES

For each day d, we employ a set of measures on the associated networks N_d^{rt}, N_d^{re}, and unigram distributions W_d^1 and W_d^2. We describe them below.

Polarization. We quantify the polarization of a network N_d by using the random-walk controversy (RWC) score introduced in previous work [12]. Intuitively, the score captures whether the network consists of two well-separated clusters.

Clustering coefficient. In an undirected graph, the clustering coefficient $cc(u)$ of a node u is defined as the fraction of closed triangles in its immediate neighborhood. Specifically, let d be the degree of node u, and T be the number of closed triangles involving u and two of its neighbors, then

$$cc(u) = \frac{2T}{d(d-1)}.$$

In our case, we consider the undirected graph G_d and compute the average clustering coefficient of all nodes that belong to each *side* – then take the mean of the two averages as the clustering coefficient of the network.

In order to control for scale effects, i.e., correlation between the size of the network (as determined by the volume of users active on day d) and the clustering coefficient, we employ a normalizer for the score. More in detail, we use an Erdős-Rényi graph as null model (with edges drawn at random among pair of nodes), and normalize the score by the expected value for a null-model graph of the same size. Unless otherwise specified, we apply the same type of normalization for all the methods defined below.

Tie strength. For each node u in a graph G_d, we consider all nodes v it is connected to across all days, and order them decreasingly by the number of occurrences $|\{d : (u, v) \in G_d\}|$. That is, the node v at the top of the list for u is the node to which v connects the most consistently throughout the time span of the dataset. Then, we define the *strong ties* of a node u as the top 10% of the nodes ordered as described above. For a given day d, we define the tie strength of a node as the number of strong ties it is connected to in the corresponding graph G_d. The tie-strength measure for the day is defined as the average tie strength for all nodes on either side. As described for the previous measure above, we normalize the reported measure by the expected value for a random graph with the same number of nodes and edges.

Cross–side openness. This measure reports the number of edges that connect nodes from opposing sides, and captures the inter-side interaction happening in the network on a given day. Formally, it is defined as

$$CSO = |\{(u, v) \in G_d : s(u) \neq s(v)\}|.$$

We apply the same normalization as described above.

Sides edge composition. For a given network, we distinguish two types of edges: *within-sides*, where both adjacent nodes belong to the same side, and *across-sides*, where the adjacent nodes belong to different sides. For each day and network, we track the fraction of the two types of edges.

Core–periphery openness. This measure is defined as the number of edges that connect a node from the core to the periphery. It captures the amount of interaction between the hard core users and the casual ones. Formally,

$$CPO = |\{(u, v) \in G_d : c(u) \wedge \neg c(v)\}|.$$

Bimotif. For a network N_d, we define the bimotif measure as the number of directed edges $(u, v) \in N_d$ for which the opposite edge (v, u) also appears in the network

$$Bimotif = |\{(u, v) \in N_d : (v, u) \in N_d\}|.$$

This measure captures the mutual interactions happening within the network. It is also known as 'reciprocity' in the literature.

Core Density. This measure captures the number of edges that connect exclusively members of the core

$$CoreDens = |\{(u, v) \in N_d : c(u) \wedge c(v)\}|.$$

Core–periphery edge composition. For a given network, we distinguish three types of edges: *core–core*, where both adjacent nodes belong to the core we have identified, *core–periphery*, where one node belongs to the core and one to the periphery, and *periphery–periphery*, where both nodes belong to the periphery. For each day and network, we track the fraction of each type of edges.

Cross–side content divergence. This measure captures the difference between the word distributions W_d^1 and W_d^2, and is based on the Jensen-Shannon divergence [22]. The Jensen-Shannon divergence is undefined when one of the two distributions is zero at a point where the other is not. Thus, we smooth the distributions by adding Laplace counts $\beta = 10^{-5}$ to avoid zero entries in either distribution.

The traffic volume on a given day can increase the vocabulary size, and thus induce an unwanted bias in the measure. In order to counter this bias, we employ a sampling procedure similar to bootstrapping from the two distributions. For each smoothed distribution W_d^1 and W_d^2, we sample with replacement $k = 10\,000$ words at random, and compute the Jensen-Shannon divergence of these equal-sized samples. We repeat the process 100 times and report the average sample Jensen-Shannon divergence as the 'cross-side content divergence' for day d. Intuitively, the higher its value, the more different the word distributions across the two sides.

Within-side entropy. This measure captures how 'concentrated' each of the two distributions W_d^1 and W_d^2 is. For each side, we compute the entropy for each distribution. The higher its value, the more widely spread is the corresponding distribution. We use the same bootstrap sampling method described above to avoid bias due to activity volume.

Topic variance. This measure captures, to some extent, *what* is being talked on the two sides of the discussion. We extract a large number of topics by using Latent Dirichlet Allocation (k=100) on the complete tweet corpus. We then compute the distribution of topics in each bucket. This distribution gives an estimate on which of the 100 topics are being talked about in the bucket. We report the variance of this distribution. If the distribution is focused on a small number of topics, the variance is high. Conversely, a low variance indicates a uniform distribution of topics.

Sentiment variance. This measure captures the variance of sentiment valence (positive versus negative) in all the tweets of one day d [12].

Psychometric analysis. To understand the if there are behavioral changes in terms of content generated and shared by users with increasing activity, we use the Linguistic Inquiry and Word Count (LIWC) dictionary,[9] which identifies emotions in words [18]. We measure the fraction of tweets containing the LIWC categories: anger, sadness, posemo, negemo, and anxiety.

4.1 Analysis

We explore how the aforementioned measures vary with the number of active users in the networks, which is a proxy for the amount of collective attention the topic attracts. We sort the time series of networks by volume of active users, and partition it into ten quantiles (each having an equal number of days), so that days of bucket i are associated with smaller volume than those of bucket j, for $i < j$. For each bucket, we report the mean and standard deviation of the values for each measure, and observe the trend from lower to higher volume.

Note that the measures presented in this section are carefully defined so that their expected value does not depend on the volume of underlying activity (i.e., number of network nodes and edges or vocabulary size).

5 FINDINGS

In what follows, we report our findings on the measures defined in Section 4 — starting from the ones related to the retweet and reply networks (Section 5.1), then proceeding to the ones related to content (Section 5.2) and network cores (Section 5.3). We provide additional analysis for the periods around the spikes in interest (Section 5.4), as well as for the evolution of measures over time (Section 5.5).

[9] http://liwc.net

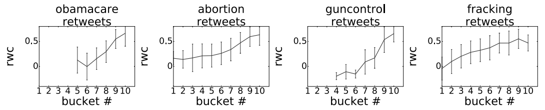


Figure 2: RWC score as a function of the activity in the retweet network. An increase in interest in the controversial topic corresponds to an increase in the controversy score of the retweet network.

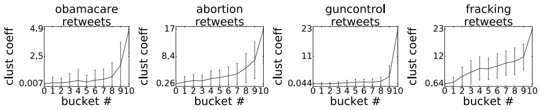


Figure 3: Average clustering coefficient of as a function of the activity in the retweet network. Spikes in interest correspond to an increase in the clustering coefficient on both sides of the discussion, which indicates the retweet networks tend to close up.

5.1 Network

We observe a significant correlation between RWC score and interest in the topic. Figure 2 shows the RWC score as a function of the quantiles of the network by retweet volume (as explained in the previous section). There is a clear increasing trend, which is consistent across topics. 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. Conversely, reply networks are sparser and more disconnected, thus, the RWC score is not meaningful in this case (not shown due to space constraints). This difference is expected, and was already observed in the work that introduced RWC [12].

A similar result can be observed for the clustering coefficient, as shown in Figure 3. As the interest in the topic increases, the two sides tend to *turtle up*, and form a more close-knit retweet network. This result suggests that the *echo chamber* phenomenon gets stronger when the discussion sparks. Our finding is also consistent with results by Romero et al. [30]. As for the previous measure, the clustering coefficient does not show a significant pattern for the reply networks. Replies are often linked to dyadic interactions, while the clustering coefficient measures triadic ones, so we expect such a difference between the two types of network.

In line with the above results, tie strength is correlated with retweet volume, as indicated by Figure 4. When the discussion intensifies, users tend to endorse the opinions of their closest friends, or their trusted sources of information. Again, this observation indicates a closing up of both sides when the debate gets heated. Interestingly, a similar trend is present for the reply network, as shown in Figure 5. Differently from previous work, we find an increase of communication of users with their strong ties, rather than with weak ties or users of the opposing side. We also observe an increase in back-and-forth communication, indicating a dialogue between users of the same side. Figure 6 shows an increase in bimotifs in the reply network when the discussion intensifies. This measure is inconclusive for the retweet network, for the reasons mentioned above.

Finally, when calculating the fractions of *within-side edges* and *across-side edges* for *across sides edge composition*, we find that reply networks typically contain higher proportions of across-side activity compared to retweet networks, consistently with earlier work. In fact, for retweet networks, almost all edges are classified as *within-side edges*. Interestingly, we also find that these proportions do not change significantly as the volume increases. The same is true for the *cross-side openness* measure (not shown).

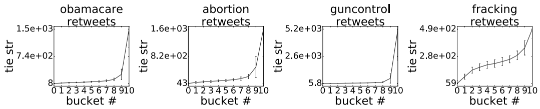


Figure 4: Tie strength as a function of the activity in the retweet network. Spikes in activity correspond to more interaction with stronger ties, which indicates a closing up of the retweet network.

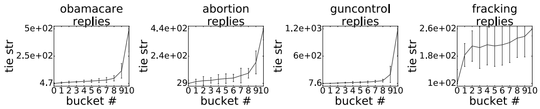


Figure 5: Tie strength as a function of the activity in the reply network. Users tend to communicate proportionally more with closer ties when interest spikes, which reveals a further closing up of the network.

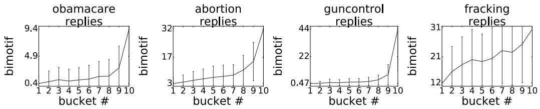


Figure 6: Bimotifs as a function of the activity in the reply network. Users tend to reciprocate the communication more as the discussion intensifies.

5.2 Content

Let us now switch our attention to the content measures. Recall that for these measures we do not distinguish between retweet and reply networks, but only between the two sides of the discussion. The main observation is that the Jensen-Shannon divergence between the two sides decreases, as shown by Figure 7. 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 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 8 shows that the entropy for one of the sides increases as interest increases

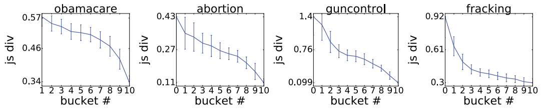


Figure 7: Jensen-Shannon divergence of the lexicon between the two sides as a function of network activity. As the interest in the topic rises, the lexicon used by the two sides tends to converge.

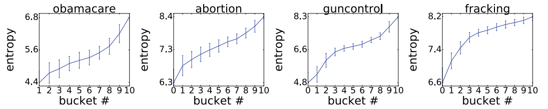


Figure 8: Entropy of the distribution over the lexicon for one side of the discussion as a function of the activity in the network (the other side shows similar patterns). As the interest increases, the entropy increases, thus indicating the use of a wider lexicon.

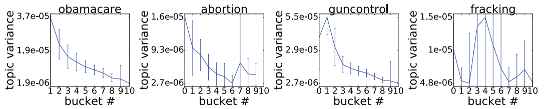


Figure 9: Variance of the topic distribution. As the interest increases, variance decreases, indicating that a wider range of topics are being discussed.

(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.

To investigate *what* causes the lexicon to be generalized, we compute the variance of the topic distribution for each bucket. As we see from Figure 9, the variance decreases with increased activity, meaning that the topic distribution becomes more uniform[10]. This result provides evidence that users do indeed discuss a wider range of topics when there is a spike in activity.

Finally, we also examine how the sentiment and other linguistic cues change with interest. We measure the variance in sentiment, fraction of tweets containing various LIWC categories, such as anger, sadness, positive and negative emotion, and anxiety. Previous work shows that sentiment variance is a measure able to separate controversial from non-controversial topics [12] and linguistic patterns of communication change during shocks [30]. However, we do not see any consistent trend. We hypothesise that this might be due to the noise in language (slang, sarcasm, short text, etc) on social media.

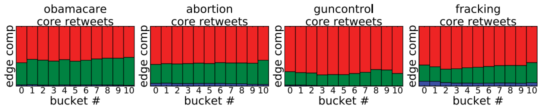


Figure 10: Edge composition as a function of network activity in the retweet network. As the interest increases, there are no major changes in the fractions of core-core (blue), core-periphery (green), and periphery-periphery (red) edges.

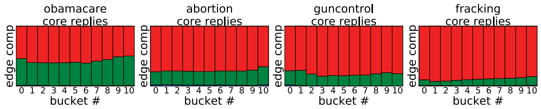


Figure 11: Edge composition as a function of network activity in the reply network. As the interest increases, there are no major changes in the fractions of core-core (blue), core-periphery (green), and periphery-periphery (red) edges.

5.3 Core

Looking at the fractions of the different types of edges (core–core, core–periphery, and periphery–periphery) across the volume buckets in Figures 10 and 11, we see that the composition of edges does not change significantly with increase collective attention. This result suggests that the discussion grows in a self-similar way.

A disproportionately large fraction of edges link the periphery to the core, when taking into account the core size, as seen in Figure 10. During a spike in interest, most casual users, who seldom participate in the discussion, endorse opinions from the core of the side they belong to (red bars). For replies, we see a similar trend with respect to activity volume in Figure 11. In general, the core is less prevalent in the discussion, as shown by the lower fraction of core-periphery edges (green bars).

However, when looking at the *core–periphery openness* (Figures 12 and 13), we see that the *normalized* number of edges between core and periphery increases, i.e., the number of edges between core and periphery increases compared to the expected number based on a random-graph null model. To interpret this result, note that when the network grows, given that the periphery is much larger than the core, most edges for the null model are among periphery nodes. Therefore, the interaction networks show a clear hierarchical structure when growing.

5.4 Local analysis

So far, we have analyzed global trends across the time series. We now focus on local trends, to drill down on what happens around the spikes, and look at local variations of the measures just before and after the spike. We mark a day in the time series as a spike if the volume of active users is at least two standard deviations above the mean. Table 2 shows the Pearson correlation between various measures and network activity, one week before and after the spike. The trends observed globally still hold. There is a positive correlation of RWC with activity, which adds more evidence to our

[10] The term ‘fracking’ is also sometimes used as an expletive, which might explain why the effects we measure are not as pronounced for this topic as the other ones. E.g. see https://twitter.com/KitKat0122/status/19820978435522561

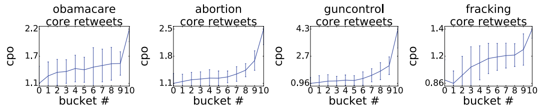


Figure 12: Core–periphery openness as a function of activity in the retweet network. As the interest increases, the number of core-periphery edges, normalized by the expected number of edges in a random network, increases. This suggests a propensity of periphery nodes to connect with the core nodes when interest increases.

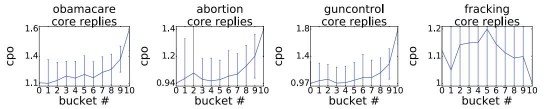


Figure 13: Core-periphery openness as a function of activity in the reply network. As the interest increases, the number of core-periphery edges, normalized by the expected number of edges in a random network, increases for most topics. This suggests a propensity of periphery nodes to connect with the core nodes when interest increases.

Table 2: Pearson correlation of various measures with volume one week before, during and after a spike in interest. All values except those marked with an asterisk (*) are significant at $p < 0.05$.

Measure	Obamacare	Gun Control	Abortion	Fracking
RWC	0.20	0.21	0.19	0.23
Openness	-0.09*	0.81	0.23	0.08
Bimotif	0.27	0.36	0.33	0.23
Tie Strength	0.96	0.98	0.95	0.86
JSD	-0.66	-0.86	-0.63	-0.46
Entropy	0.42	0.46	0.67	0.26
Frac. RT	0.15*	0.6	0.59	0.56
Frac. Men.	0.20	0.71	0.54	0.51
Frac. URL	0.32	0.36	0.39	0.40

finding that polarization increases during spikes. The trends for bimotif, tie strength, and content divergence also persist, and are much stronger locally.

In addition to the previous measures, we also analyze other content features, such as the fraction of retweets, replies, mentions, and URLs around the spike. Interestingly, we find strong positive correlation of retweets, mentions, and URLs with volume, which indicates that discussion and endorsement increase during a spike. This finding is consistent with the ones by Smith et al. [31], who find that users tend to add URLs to their tweets when discussing controversial topics. Note that these additional content measures are only indicative for the local analysis, and do not produce consistent results at the global level.

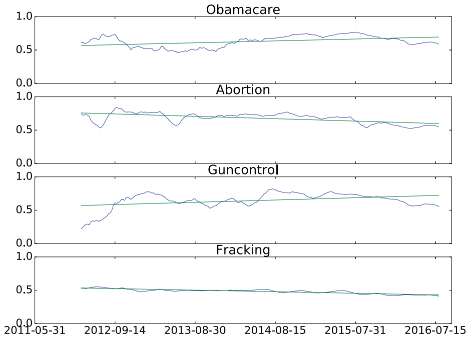


Figure 14: Long-term trends of RWC (controversy) score in our dataset. No consistent trend can be observed, which contradicts the narrative that social media is making our society more divided.

5.5 Evolution over time

Let us now focus on how the measures change throughout time. The longitudinal span of the dataset of five years allows us to track the long-term evolution of discussion on controversial topics. A common point of view holds that social media is aggravating the polarization of society and exacerbating the divisions in it [5]. At the same time, the political debate (in U.S.) itself has become more polarized in recent years [4]. However, we do not find conclusive evidence for this argument with our analysis on this dataset.

Figure 14 shows the long-term trends of the RWC measure for the four topics. The trend is downwards for 'abortion' and 'fracking', while it is upwards for 'obamacare' and 'gun control'. One could argue that the latter topics are more politically linked to the current administration in U.S., and for this reason have received increasing attention with the elections approaching. However, the only safe conclusion that can be drawn from this dataset is that there is no clear signal. The figure suggests that social media, and in particular Twitter, are better suited at capturing the 'twitch' response of the public to events and news. In addition, while our dataset spans a quite long time span for typical social media studies, it is still much shorter than other ones used typically in social science (coming from, e.g., census, polls, congress votes). This limit is intrinsic of the tool, given that social media have risen in popularity only relatively recently (e.g., Twitter is 10 years old).

5.6 Non-controversial topics

For comparison, we perform measurements over a set of non-controversial topics, defined by the hashtags *#ff*, standing for 'Follow Friday', used every Friday by users to recommend interesting accounts to follow; *#nba* and *#nfl*, used to discuss sports games; *#sxsw*, used to comment on the *South-by-South-West* conference; *#tbt*, standing for 'Throwback Thursday', used every Thursday by users to share memories (news, pictures, stories) from the past.

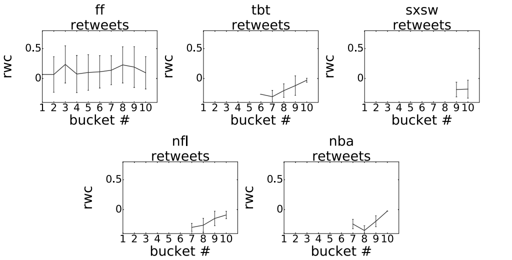


Figure 15: Non-controversial topics: RWC score as a function of the activity in the retweet network.

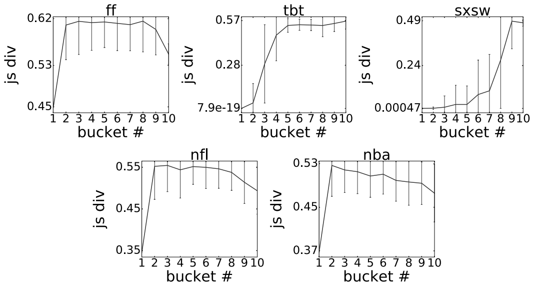


Figure 16: Non-controversial topics: Jensen-Shannon divergence of the lexicon between the two sides as a function of network activity. As the interest in the topic rises, the lexicon used by the two sides tends to converge.

We find that several structural measures, namely *clustering coefficient*, *tie strength*, and *bimotif*, behave similarly to the controversial topics, in that they obtain increased values for increased volume of activity. This result is in accordance with the ones by Romero et al. [30]. Conversely, the values of the RWC measure typically remain in ranges that indicate low presence of controversy, even as the volume of activity spikes (Figure 15). Additionally, with the definition of 'core' introduced above, we could only identify a negligibly small core for these topics (i.e., found very few users who were consistently active on these topics).

Finally, in terms of content measures we find that, as for the controversial topics, the entropy of the lexicon increases with volume (Figure 17). Topic variance also decreases with volume in most cases, meaning that a wider range of topics are discussed (Figure 18). On the contrary, the Jensen-Shannon divergence stays at relatively constant values across volume levels (Figure 16). It thus behaves differently compared to controversial topics (Figure 7). This result is to be expected, as the two 'sides' identified by METIS on the networks of non-controversial topics are not as well defined as they are in the case of controversial topics.

6 CONCLUSION

The evolution of networks is a well-studied phenomenon in social sciences, physics, and computer science. However, the evolution of *interaction* networks has received substantially less attention so far. In particular, interaction networks related to discussions of controversial topics, which are important from a sociological point of view, have not been analyzed before. This study is a first step towards understanding this important social phenomenon.

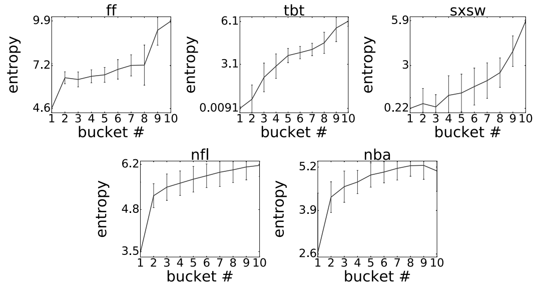


Figure 17: Non-controversial topics: Entropy of the distribution over the lexicon for one side of the discussion as a function of the activity in the network (the other side shows similar patterns).

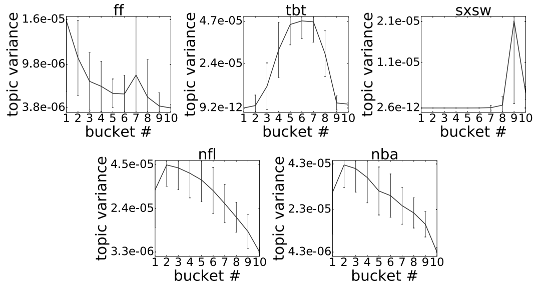


Figure 18: Non-controversial topics: Variance of the topic distribution. As the interest increases, variance decreases, indicating that a wider range of topics are being discussed.

We analyzed four highly controversial topics of discussion on Twitter for a period of five years. By examining the endorsement and communication networks of users involved in these discussions, we found that spikes in interest correspond to an increase in the controversy of the discussion. This result is supported by a wide array of network analysis measures, and is consistent across topics. We also found that interest spikes correspond to a convergence of the lexicon used by the opposite sides of a controversy, and a more uniform lexicon overall. The code and datasets used in the paper are available on the project website.[8]

Implications of this work relate to the understanding of how our society evolves via continuous debates, and how culture wars develop [1, 16, 23]. It is often argued that technology, and social media in particular, is having a negative impact on our ability to relate to the unfamiliar [5], due to the "echo chamber" and "filter bubble" effects. However, while we found instantaneous temporary increase in controversy in relation to external events, our study did not find evidence of long term increase in polarization of the discussions, neither after these events nor as a general longitudinal trend. At the same time, investigating how to reduce the polarization of these discussions on controversial topics is a research-worthy problem [11, 13], and taking into account the dynamics of the process is a promising direction to explore.

Our observations pave the way to the development of models of evolution for controversial interaction networks, similarly to how studies about measuring the Web and social media were the stepping stone to developing models for them. A logical next step for this line of work is to investigate how to use early signals from social media network structure and content to predict the impact of an event. Equally of interest is whether the observations made in this study translate to other social media beside Twitter, for instance, Facebook or Reddit. Finally, while we did not find any consistent long-term trend in the polarization of the discussions, it is worth continuing this line of investigation, as the effects of increased polarization might not be easily discoverable from social-media analysis alone.

Acknowledgements. This work has been supported by the Academy of Finland project "Nestor" (286211) and the EC H2020 RIA project "SoBigData" (654024).

7 REFERENCES

[1] Alan Abramowitz and Kyle Saunders. 2005. Why can't we all just get along? The reality of a polarized America. In *The Forum*, Vol. 3. bepress, 1–22.

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

[3] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. 2007. Analysis of topological characteristics of huge online social networking services. In *WWW*. ACM, 835–844.

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

[5] Yochai Benkler. 2006. *The wealth of networks: How social production transforms markets and freedom.* Yale University Press.

[6] Alessandro Bessi, Guido Caldarelli, Michela Del Vicario, Antonio Scala, and Walter Quattrociocchi. 2014. Social determinants of content selection in the age of (mis) information. In *SocInfo*. Springer, 259–268.

[7] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. In *ICWSM*.

[8] Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. 2012. From Chatter to Headlines: Harnessing the Real-Time Web for Personalized News Recommendation. In *WSDM*. 153–162.

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

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

[11] Kiran Garimella, Gianmarco De Francisc iMorales, Aristides Gionis, and Michael Mathioudakis. 2017. Mary, Mary, Quite Contrary: Exposing Twitter Users to Contrarian News. In *WWW*. 201–205.

[12] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying Controversy in Social Media. In *WSDM*. ACM, 33–42.

[13] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. Reducing Controversy by Connecting Opposing Views. In *WSDM*. ACM, 81–90.

[14] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. The Ebb and Flow of Controversial Debates on Social Media. In *ICWSM*.

[15] Kiran Garimella and Ingmar Weber. 2017. A Long-Term Analysis of Polarization on Twitter. In *ICWSM*.

[16] Benjamin Highton and Cindy D Kam. 2011. The long-term dynamics of partisanship and issue orientations. *The Journal of Politics* 73, 01 (2011), 202–215.

[17] George Karypis and Vipin Kumar. 1995. METIS - Unstructured Graph Partitioning and Sparse Matrix Ordering System. (1995).

[18] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790.

[19] Janette Lehmann, Bruno Gonçalves, José J. Ramasco, and Ciro Cattuto. 2012. Dynamical Classes of Collective Attention in Twitter. In *WWW*. ACM, 251–260.

[20] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. 2008. Microscopic evolution of social networks. In *KDD*. ACM, 462–470.

[21] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2005. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD*. ACM, 177–187.

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

[23] Kara Lindaman and Donald P Haider-Markel. 2002. Issue evolution, political parties, and the culture wars. *Political Research Quarterly* 55, 1 (2002), 91–110.

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

[25] Haokai Lu, James Caverlee, and Wei Niu. 2015. BiasWatch: A Lightweight System for Discovering and Tracking Topic-Sensitive Opinion Bias in Social Media. In *CIKM*. ACM, 213–222.

[26] Yelena Mejova, Amy X Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and Sentiment in Online News. *Symposium on Computation + Journalism* (2014).

[27] AJ Morales, J Borondo, JC Losada, and RM Benito. 2015. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos* 25, 3 (2015).

[28] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *arXiv:1306.5204* (2013).

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

[30] Daniel M Romero, Brian Uzzi, and Jon Kleinberg. 2016. Social Networks Under Stress. In *WWW*. 9–20.

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

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

Publication 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.

Factors in Recommending Contrarian Content on Social Media

Kiran Garimella
Aalto University
Helsinki, Finland
kiran.garimella@aalto.fi

Gianmarco De Francisci Morales
Qatar Computing Research Institute
Doha, Qatar
gdfm@acm.org

Aristides Gionis
Aalto University
Helsinki, Finland
aristides.gionis@aalto.fi

Michael Mathioudakis
Aalto University
Helsinki, Finland
michael.mathioudakis@aalto.fi

ABSTRACT

Polarization is a troubling phenomenon that can lead to societal divisions and hurt the democratic process. It is therefore important to develop methods to reduce it.

We propose an algorithmic solution to the problem of reducing polarization. The core idea is to expose users to content that challenges their point of view, with the hope broadening their perspective, and thus reduce their polarity. Our method takes into account several aspects of the problem, such as the estimated polarity of the user, the probability of accepting the recommendation, the polarity of the content, and popularity of the content being recommended.

We evaluate our recommendations via a large-scale user study on Twitter users that were actively involved in the discussion of the US elections results. Results shows that, in most cases, the factors taken into account in the recommendation affect the users as expected, and thus capture the essential features of the problem.

1 INTRODUCTION

Polarization around controversial issues is a well-studied phenomenon in the social sciences [11].Social media have arguably amplified polarization, thanks to the scale of discussions and their publicity [7]. This paper studies how to reduce polarization on social media by recommending *contrarian* content, i.e., content that expresses a point-of-view opposing the one held by the target user. In particular, we examine which features might be used to develop such a content recommender system.

We focus on controversial issues that create discussions online. Usually, these discussions involve a fair share of “retweeting” or “sharing” opinions of authoritative figures with whom the user agrees. Therefore, it is natural to model the discussion as an *endorsement graph*: a vertex u represents a user, and a directed edge (u, v) represents the fact that user u endorses the opinion of user v.

Due to phenomena such as homophily, confirmation bias, and selective exposure, social media often create echo chambers [5, 8]. These chambers, in turn, cultivate isolation and misunderstanding in society [18], and deepen its polarization.

A potential solution to this problem is to encourage users to consider points of view different from their own. Thus, in this paper, we study methods to recommend content items (e.g., news articles, opinion pieces, blog posts) that express a contrarian point of view, while at the same time being appealing to the target user.

In particular, given metrics that measure the polarization of users and items (such as those proposed in recent research [3]), our goal is to recommend an item that nudges the user towards the opposite polarity. That is, we seek to propose content produced by a user v to another user u, thus informing u of a different viewpoint, and hoping that u will endorse v.

Clearly, some content is more likely to be endorsed than other. For instance, people in the “center” might be easier to convince than people on the two extreme ends of the political spectrum [13]. We take this issue into account by modeling the *acceptance probability* for a recommendation as a separate component of the model.

We blend these factors, together with other signals such as topic and popularity, to create a ranked list of recommendations. Our solution employs a well-known weighted rank-aggregation algorithm at its core [17].

We evaluate our proposal by running an online user study with Twitter users. We focus on the recent 2016 US presidential elections, and generate recommendations for the thousands of users involved in this highly-polarizing controversial discussion. The results of the study show that the two main factors used in the recommendation, the polarity and the acceptance probability models, are predictive of the responses of the users.

In summary, we make the following contributions:

- We study the problem of bridging echo chambers algorithmically, in a language- and domain-agnostic way. Previous studies that address this problem focus mostly on understanding *how* to recommend content to an ideologically opposite side, while we focus on *which* contrarian content to recommend. We believe that the two approaches complement each other in bringing us closer to bursting filter bubbles.

WebSci’17, June 25-28, 2017, Troy, NY, USA.

978-1-4503-4896-6/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3091478.3091515

- We build on top of results from recent user studies [14, 15, 19] on how users prefer to consume content from opposing views, and formulate the task as a content-recommendation problem based on an endorsement graph, while also taking into account the acceptance probability of a recommendation.
- We evaluate the proposed solution via a user study on Twitter users, and demonstrate the validity of the main factors involved in the recommendation.

2 RELATED WORK

Although the Web was envisioned as a place of open discussions on a wide range of topics, many people tend to restrict themselves to viewing and sharing information that conforms with their beliefs. A wide body of recent studies has explored [1, 2] and quantified [3] the notions of "filter bubble" and "echo chambers".

Munson et al. [15] created a browser widget that measures the bias of users based on the news articles they read. Their study shows that users are willing to slightly change views once they are shown their biases. Graells-Garrido et al. [9] 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. Graells-Garrido et al. [10] propose to find 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. Park et al. [16] propose methods for presenting multiple aspects of news to reduce bias.

Most relevant to this work is the recent study about the problem of reducing the overall polarization of a controversial topic in a network [6]. The study tries to find the best edges that can be added to an endorsement graph so that the polarization score of the network is reduced. In this paper, we focus on reducing the polarization of an individual user (local objective), instead of the entire network (global objective).

There have also been a number of demos and systems: Wall Street Journal's *Blue feed-Red feed*[1] raises awareness about the extent to which viewpoints on a matter can differ, by showing side-by-side articles expressing very liberal and very conservative viewpoints; *Politecho*[2] displays how polarizing the content on a user's news feed is when compared to their friends'; *Escape your bubble*[3] is a browser extension to add hand-curated content from the opposite side in Facebook; automated bots have been created to respond to posts containing harassment or fake news,[4] with an attempt to de-polarize the discussion and educate users. Moreover, new social media platforms have been proposed that aim to be designed in such a way to encourage discussions and debates, such as the Filterburst project,[5] Rbutr,[6] where users can post rebuttals of other urls, and a wikipedia for debates.[7]

The proposed method differs from existing ones in many ways. First, our approach is completely algorithmic, unlike most demos listed above, which involve manual curation. Second, as discussed above, it builds on top of existing research and incorporates key findings of previous work.

3 PRELIMINARIES

A topic of discussion is identified as the set of tweets that satisfy a text query – e.g., all tweets that contain a specific hashtag. We represent a topic with an *endorsement graph* $G(V, E)$, where vertices V represent users and edges E represent *endorsements*.

It has been shown that an endorsement graph captures well the extent to which a topic is controversial [3]. In particular, the endorsement graph of a controversial topic has a *multimodal clustered structure*, where each cluster of vertices represents one viewpoint on the topic. As we focus on two-sided controversies, we identify the two sides of a controversial topic by employing a *graph-partitioning* algorithm, which partitions the graph into *two* subgraphs. In this work, we specifically focus on recommending content in the form of news items, such as articles, blog posts, and opinion pieces. The item pool for the recommendation comprises all the links shared by the active users during the observation window.

User polarization score. We use a recently-proposed methodology to define the polarization score for each user in the graph [4]. The score is based on the expected hitting time 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. 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^+ 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

$$\rho(u) = \rho^X(u) - \rho^Y(u) \quad \in (-1, 1). \tag{1}$$

Following this definition, 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 $\rho(u) \approx 1 - 0 = 1$. The opposite is true for vertices that are far from X^+ but close to Y^+; leading to a polarization score $\rho(u) \approx -1$.

Item polarization score. Once we have obtained polarization scores for users in the graph, it is straightforward to derive a similar score for content items shared by these users. Specifically, we define the polarization score of an item i as the average of the polarization scores of the set of users who have shared i, denoted by U_i:

$$\rho(i) = \frac{1}{|U_i|} \sum_{u \in U_i} \rho(u) \quad \in (-1, 1). \tag{2}$$

Acceptance probability. Not all recommendations are agreeable, especially if they do not conform to the user's beliefs. To reduce

[1] http://graphics.wsj.com/blue-feed-red-feed/
[2] http://politecho.org/
[3] https://www.escapeyourbubble.com/
[4] http://wpo.st/4kVR2, https://goo.gl/Xl6x9t
[5] http://www.filterburst.com/
[6] http://rbutr.com/
[7] http://www.debatepedia.org/en/index.php/Welcome_to_Debatepedia%21

these effects, we define an acceptance probability, which quantifies the degree to which a user is likely to endorse the recommended content. We use the item and user polarization scores defined above to estimate the likelihood that a target user u endorses (i.e., retweets) the recommended item i. We build an acceptance model by adapting a similar one based on the feature of user polarization [6]. High absolute values of user polarization (close to -1 or $+1$) indicate that the user belongs clearly to one side of the controversy, while middle-range values (close to 0) indicate that the user is in the middle of the two sides. It was shown that users from either side accept content from different sides with different probabilities, and these probabilities can be inferred from the graph structure [6]. For example, a user with polarization close to -1 is more likely to endorse a user with a negative polarization than a user with polarization $+1$. This intuition directly translates to endorsing items, and therefore can be used for our recommendation problem.

Based on this intuition, we define the acceptance probability $p(u,i)$ of a user u endorsing item i as

$$p(u,i) = {}^{N_e(\rho(u),\,\rho(i))}/_{N_x(\rho(u),\,\rho(i))}, \tag{3}$$

where $N_e(\rho(u),\rho(i))$ and $N_x(\rho(u),\rho(i))$ are the number of times a user with polarity $\rho(u)$ has endorsed or was exposed to (respectively) content of polarity $\rho(i)$. In practice, the polarity scores are bucketed to smooth the probabilities.

4 RECOMMENDATION FACTORS

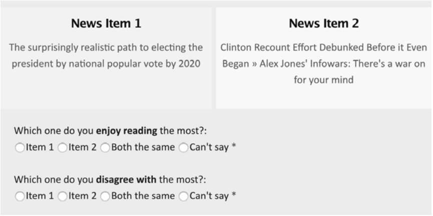
This section describes the factors used to generate recommendations. Though our main focus is to connect users with content that expresses a contrarian point of view, we also want to maximize the chances of such a recommendation being endorsed by the user. We take into account several factors: reduction in polarization score of the target user; exclusivity of the candidate items (polarity of the items); acceptance probability of recommendation based on polarization scores; topic diversity; popularity/quality of the candidate item. Next, we describe these factors in more detail.

Reduction of user polarization score. The maximum reduction of user polarization score is achieved by putting the user in contact with an authoritative source from the opposing side. Leveraging this idea, we build a list of items L_1 by considering items shared by high degree nodes on the opposite side of the target user, and ranking them by the potential decrease in user polarization score.

Exclusivity on either side. We consider items that are almost exclusively shared by one of the sides. Specifically, we denote by n_i^X and n_i^Y the number of users who shared each item i on side X and Y, respectively. For each side, we generate a list L_2 ranked by the ratio of shares ${}^{n_i^X}/_{n_i^Y}$ (for side X) and ${}^{n_i^Y}/_{n_i^X}$ (for side Y).

Acceptance probability. For a given user u, all items sorted in decreasing order of acceptance probability $p(u,i)$ make up list L_3.

Topic diversity. We want to ensure that the recommendations are topically diverse. 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 item. Topics are defined as *named entity*, and we extract them using the tool tagme.[8] Given a user vector t_u, we compute the cosine similarity with all item vectors t_i, and rank items in increasing order of cosine similarity (list L_4).



News Item 1

The surprisingly realistic path to electing the president by national popular vote by 2020

News Item 2

Clinton Recount Effort Debunked Before it Even Began » Alex Jones' Infowars: There's a war on for your mind

Which one do you **enjoy reading** the most?:

○Item 1 ○Item 2 ○Both the same ○Can't say *

Which one do you **disagree with** the most?:

○Item 1 ○Item 2 ○Both the same ○Can't say *

Figure 1: Screenshot of the interface shown for a user with a high polarity on the political left (Democrat).

Popularity on either side. Finally, we 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. We produce list L_5 of items in decreasing popularity score.

Rank Aggregation. Given the 5 ranked lists discussed above, we use a weighted rank-aggregation scheme to generate the final recommendations. The intuition behind using rank aggregation is that items that are highly ranked in many lists, are also highly ranked in the output list. In particular, we use a weighted rank-aggregation technique proposed by Pihur et al. [17], whose goal is to minimize the objective function

$$\phi(\delta) = \sum_{i=1}^{5} w_i d(\delta, L_i), \tag{4}$$

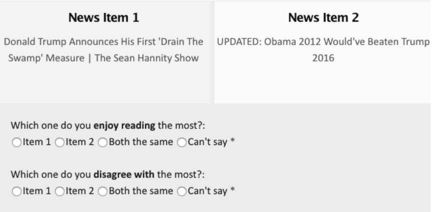
where δ is the optimal ranked output list, d is any distance function (we use the Spearman footrule distance), and w_i are the importance weights of each list. We can set the weights to generate highly contrarian recommendations (by giving large weights to L_1 and L_2) or recommendations that are likely to be accepted (by giving large weight to L_3).

5 EVALUATION

Dataset. We collect all tweets containing the hashtag `#USelections`, used in discussions about the US presidential elections during Nov 9–12, 2016. From the 6.2 M tweets collected, we build an endorsement graph with 6764 nodes (users) and 9896 edges (retweets). To filter out noise, the graph contains an edge between two users only if at least 5 retweets between the two users occur. We partition the graph to obtain the two sides by using METIS [12]. For recommendation items (urls), we consider items that have been shared at least 5 times in our dataset. The final pool contains 10 210 candidate items, which include news articles, blog posts, opinion pieces, etc.

User study. We run an online user study involving all 6764 users in the dataset with the aim of evaluating how users percieve the two main conflicting factors proposed, i.e. the contrarian features (L_1, L_2) and acceptance features (L_3). 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. In more

[8] https://services.d4science.org/web/tagme



News Item 1

Donald Trump Announces His First 'Drain The Swamp' Measure | The Sean Hannity Show

News Item 2

UPDATED: Obama 2012 Would've Beaten Trump 2016

Which one do you **enjoy reading** the most?:

○Item 1 ○Item 2 ○Both the same ○Can't say *

Which one do you **disagree with** the most?:

○Item 1 ○Item 2 ○Both the same ○Can't say *

Figure 2: Screenshot of the interface shown for a user with a high polarity on the political right (Republican).

Table 1: Results from the user study.

Main factor	Item1 (Acceptance)	Item2 (Contrarian)	Both the same	Can't say
Enjoy	51	19	8	15
Disagree	22	57	7	7

detail, by using the methodology described above, we compute two recommendations for each user: in the first one we give a high weight (60%) to contrarian features (L_1 and L_2), while in the second one we give high weight (60%) to acceptance probability (L_3). We distribute the remaining 40% equally among other features.

The main research questions we investigate are: (*i*) is a high acceptance probability factor predictive of content with higher acceptance? and (*ii*) are contrarian factors predictive of more disagreement with the user? To simplify the task for the user, we set up the user study as a relative comparison between the two recommendations, rather than asking for absolute judgments. Since the two recommendations are generated completely independently, we assume that they do not influence the users decision making process in choosing one over the other.

We create a web form[9] with two recommended items, customized for each user, with the item weighted by the acceptance features shown on the left and contrarian features on the right. Figures 1 and 2 show two instances of the web form. Looking at Figure 1, given the left-leaning political affiliation of the user, the recommendation on the left side (News item 1) looks more agreeable than the recommendation on the right side (News item 2). The opposite is true for Figure 2, which targets a right-leaning user.

We contacted users on Twitter with the following private message: "@username We are scientists studying social media. Would u like to help science by participating in a survey? http://bit.ly/XXXXX'', and waited for two weeks for them to respond. In total, we sent around 6700 messages and received 93 valid responses after removing duplicates (1.4% response rate).

Our expectation is that users enjoy reading the item with high acceptance probability, and disagree with the contrarian item. The results, summarized in Table 1, confirm our expectations. Indeed, most users enjoy reading the item with high acceptance, and disagree with the contrarian item. Specifically, 44 out of the 93 users (47%) reported that at the same time they enjoy the first item, and disagree with the second. For a few users (n=7), we were able to generate enjoyable recommendations that they disagreed with. While this was not the goal of the specific user study, it is indeed our ultimate goal, and thus these results are highly encouraging.

Acknowledgements. This work has been supported by the Academy of Finland project "Nestor" (286211) and the EC H2020 RIA project "SoBigData" (654024).

[9] http://bit.ly/2jOQBxP

6 REFERENCES

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

[2] Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. In *ICWSM*.

[3] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying Controversy in Social Media. In *WSDM*. 33–42.

[4] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying Controversy in Social Media. *arXiv preprint arXiv:1507.05224* (2016).

[5] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. The Ebb and Flow of Controversial Debates on Social Media. In *ICWSM*.

[6] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. Reducing Controversy by Connecting Opposing Views. In *WSDM*. 81–90.

[7] Kiran Garimella and Ingmar Weber. 2017. A Long-Term Analysis of Polarization on Twitter. In *ICWSM*.

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

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

[10] Eduardo Graells-Garrido, Mounia Lalmas, and Daniele Quercia. 2014. People of opposing views can share common interests. In *WWW Companion*. 281–282.

[11] Daniel J Isenberg. 1986. Group polarization: A critical review and meta-analysis. *Journal of personality and social psychology* (1986).

[12] George Karypis and Vipin Kumar. 1995. METIS - Unstructured Graph Partitioning and Sparse Matrix Ordering System. (1995).

[13] Q Vera Liao and Wai-Tat Fu. 2014. Can you hear me now?: mitigating the echo chamber effect by source position indicators. In *CSCW*. 184–196.

[14] Q Vera Liao and Wai-Tat Fu. 2014. Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions. In *CHI*. 2745–2754.

[15] Sean A Munson and others. 2013. Encouraging Reading of Diverse Political Viewpoints with a Browser Widget.. In *ICWSM*.

[16] Souneil Park and others. 2009. NewsCube: delivering multiple aspects of news to mitigate media bias. In *CHI*. 443–452.

[17] Vasyl Pihur and others. 2009. RankAggreg, an R package for weighted rank aggregation. *BMC bioinformatics* 10, 1 (2009), 1.

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

[19] VG Vydiswaran, ChengXiang Zhai, Dan Roth, and Peter Pirolli. 2015. Overcoming bias to learn about controversial topics. *JASIST* (2015).

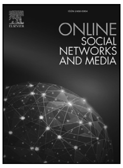
Publication VII

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.

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem



Automatic controversy detection in social media: A content-independent motif-based approach

Mauro Coletto[a,b,*], Kiran Garimella[c], Aristides Gionis[c], Claudio Lucchese[a,b]

[a] Ca' Foscari University, Venice, Italy
[b] ISTI-CNR, Pisa, Italy
[c] Aalto University, Helsinki, Finland

ARTICLE INFO

Article history:
Received 23 August 2017
Revised 12 October 2017
Accepted 13 October 2017

Keywords:
Controversy detection
Polarization
Social network analysis
Twitter
Motif
Social media

ABSTRACT

Online social networks are becoming the primary medium by which people get informed, as they provide a forum for expressing ideas, contributing to public debates, and participating in opinion-formation processes. Among the topics discussed in Social Media, some lead to controversy.

Identifying controversial topics is useful for exploring the space of public discourse and understanding the issues of current interest. Thus, a number of recent studies have focused on the problem of identifying controversy in social media mostly based on the analysis of textual content or rely on global network structure. Such approaches have strong limitations due to the difficulty of understanding natural language, especially in short texts, and of investigating the global network structure.

In this work, we show that it is possible to detect controversy in social media by exploiting network motifs, i.e., local patterns of user interaction. The proposed approach allows for a language-independent and fine-grained analysis of user discussions and their evolution over time. Network motifs can be easily extracted both from user interactions and from the underlying social network, and they are conceptually simple to define and very efficient to compute. We assess the predictive power of motifs on a manually labeled twitter dataset. In fact, a supervised model exploiting motif patterns can achieve 85% accuracy, with an improvement of 7% compared to baseline structural, propagation-based and temporal network features. Finally, thanks to the locality of motif patterns, we show that it is possible to monitor the evolution of controversy in a conversation over time thus discovering changes in user opinion.

1. Introduction

The usage of online social networks is becoming an increasing trend through which people around the globe are in contact with others and get informed about topics of interest. Additionally, online social networks provide a forum for expressing ideas, contributing to public debates, and participating in opinion-formation processes. Even though many studies have been devoted to understand different aspects of social network structure and function, such as, community structure [1], information spreading [2], information seeking [3], link prediction [4], etc., much less work is available on analyzing online discussions and public debates.

In this paper, we study the problem of identifying controversies in social media, one of the many different aspects of analyzing online discussions and understanding how people participate in those. The problem of studying controversy in social media has recently drawn some attention [5,6]. However, as this is a difficult problem, involving processing of human language and network dynamics, existing studies have limitations. For example, many papers study controversy in very controlled case studies, or focus on a predefined topic, most typically politics [7], for which they employ auxiliary domain-specific sources and datasets. In other cases, proposed approaches are based on content-based analysis [8], which has several limitations, as well, due to the ambiguity of the language and the fact that models become language-dependent and topic-dependent.

Instead, in this paper we aim to identify controversies on *any* topic, discussed in *any* language. Given this objective, our approach is based on the analysis of the *network structure*. In this sense, our paper is related to the recent work of Garimella et al. [5], who also aim at identifying controversies in the wild, independent of topic or language. In that work, the authors focus on a topic defined by a single hashtag, and then analyze the retweet network after partitioning it into two clusters (the two sides of controversy).

[*] Corresponding author.
E-mail addresses: mauro.coletto@imtlucca.it, mauro.coletto@unive.it, mauro.coletto@isti.cnr.it (M. Coletto), kiran.garimella@aalto.fi (K. Garimella), aristides.gionis@aalto.fi (A. Gionis), claudio.lucchese@unive.it, claudio.lucchese@isti.cnr.it (C. Lucchese).

https://doi.org/10.1016/j.osnem.2017.10.001
2468-6964/

An obvious limitation in their work is that they assume that a topic partitions the network into two clusters (while none, or more than two clusters, may be present), and that it is computationally feasible to identify those clusters. In our work, we overcome those limitations by analyzing local network patterns (*motifs*), and thus, making no assumption about the global cluster structure of the network, or about our ability to detect network clusters. Moreover, note that the separation of the retweet network in communities does not always reflect controversy; it may also mean that a hashtag is used in two communities with different acceptations. Our model catches antagonism in the conversation and, in fact, we find that some hashtags (#germanwings, #onedirection) that were detected as not controversial by previous studies, contain controversial discussions. Finally, in the work of Garimella et al. [5] the approach of detecting controversy is static and is based on analyzing the retweets of a given hashtag. In our case we focus on the analysis of the discussions generated by those tweets. This allows us to discover potentially controversial sub-topics that may be present within an otherwise non-controversial topic.

We propose the use of motifs extracted from the user reply and friendships graphs to detect controversial threads of discussion in online social networks. The proposed motifs can be easily computed as they encompass interactions among two or three users only. Being graph-based, such motifs are language independent and topic independent: they can be applied to investigate interactions in social networks without any additional domain knowledge. We measure the predictive power of the proposed motifs on a collection of Twitter data. We found that local motifs can improve the accuracy of frequently used graph-based features (e.g., cascade depth, inter-reply time) achieving an accuracy of 85%. We claim that such motifs are able to model both user homophily, through the friendship graph, and user interest in discussing specific topics even beyond their social circles, through the reply graph.

This paper is an extended version of a previous conference paper [9]. The original contributions presented in this paper include: a more detailed description of the proposed method, a dynamic use of the method, an additional experiment on a dataset based on Twitter hashtags (Dataset2: *Twitter hashtags.*). We applied the method to specific accounts (Dataset1: *Twitter pages.*), but also to specific concepts, represented by Twitter hashtags. We used a previously used Twitter hashtags dataset in order to compare our approach to previous ones and we report the analyses. Finally, the proposed motifs, being local to two or three users, allow a fine-grained analysis of the evolution of a discussion over time and of the interactions among its users. We extended the conference paper with the description of a temporal variant of the method, reporting some relevant examples. In fact, we found that non controversial conversations happen to become controversial either limitedly to a sub-tree of the discussion thread, or globally due for instance to external events such as news.

2. Related work

Controversy and polarization. The analysis of controversy on the web and social media has received considerable attention in recent years, with a number of papers studying controversy on general web pages [10], blogs [11], online news [8,12], and social media [5,7,13].

The existence of polarization on social media was first studied by Adamic and Glance [11] who identified a clear separation in the hyperlink structure of political blogs. Conover et al. [7] studied this phenomenon on Twitter, evaluating the polarization on the retweet network. In a more recent work, Garimella et al. [5] showed that the polarized structure in the retweet graph extends beyond politics. They also proposed algorithmic methods to measure the amount of controversy on a topic, by considering the structure of the network formed by retweets and followers. In a similar spirit, Guerra et al. [14] considered a measure based on boundary connectivity patterns in order to identify if a discussion is controversial. Other approaches have also been proposed to identify controversy on social media at a *user* level. For example, BiasWatch is a weakly-supervised approach fusing content and network data to infer user polarity [6].

Controversies are inherently dynamic. Non-controversial topics could become controversial and vice-versa. Morales et al. [15] present an approach based on label propagation in order to quantify the level of controversy in the network. They apply their measure on Twitter data from Venezuela over a long period and showed that they can capture real-life shifts in polarization. Coletto et al. [16] proposed an approach for jointly tracking user polarity and topic evolution. The method proposed in this paper can handle the dynamic nature of a controversial topic.

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 conversation graphs on Twitter [17,18]. Those studies analyze various types of conversation graphs, such as *long path-like reply trees, large star-like trees*, and *long irregular trees*. They also show that paths are making up to 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 controversial discussions, we additionally detect long trees with multiple branches indicating the different threads of the discussions, e.g., see Fig. 1.

Analysis of conversation graphs in rumor and misinformation spreading has shown that information flow in the network gives rise to certain types of local patterns [19,20]. Smith et al. [21] study the role of social media in the discussion of controversial topics. They try to understand reply and retweet interactions at a user level and conclude that We that users are quicker to spread information that agrees with their position more often.

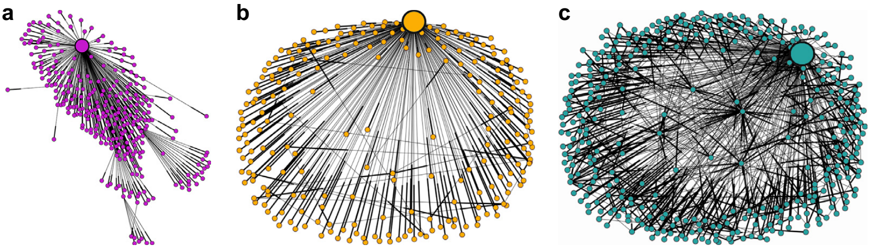


Fig. 1. Examples of different user-interaction networks: (a) content reply tree; (b) user reply graph for a non-controversial conversation; and (c) user reply graph for a controversial conversation.

However, to our knowledge, this is the first work to do an in-depth study of the role of network motifs in the context of identifying controversy in social media.

Motifs indicate patterns of interactions/interconnections in complex networks. The work of Milo et al. [22] was one of the first to analyze the occurrence of different motifs in networks arising in a wide range of fields, from biochemistry to engineering. Their finding that "*motifs may thus define universal classes of networks*" is one of our motivations for exploring simple interaction patterns related to controversy.

In the context of social networks, motifs may indicate a specific function or role of certain nodes. For example, network motifs have been used recently to explain higher-order network organization, and subsequently, use this information to cluster networks [23].

Conversation textual analysis. The problem of detecting disagreement in conversation text was recently studied by Allen et al. [24], who use rhetorical structure features to identify disagreement. They claim that this is a difficult task, even for humans.

Most related to our paper is the work by Chen and Berger [25], who 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.

Furthermore, language-analysis tools have been used widely to determine the emotional tone of a conversation [26], e.g., whether a message is partial or impartial [27], subjective/objective, positive/negative [28], etc.

All the different methods discussed above use only textual information. Even though the use of text features is orthogonal to our method, and they can be added separately, we chose not to do so explicitly, since text-analysis tools are language dependent, and since we are mainly interested in contrasting network motifs with other network-structure features.

3. Data collection

We consider two Twitter datasets.

Dataset1: ***Twitter pages.*** Our main source of data is a carefully-curated set of popular Twitter pages which covers a wide range of domains (news, politics, celebrity, gossip, entertainment) and languages. The way we choose popular pages is generic and can be emulated on other social networks. For each page, we gather the last two hundreds tweets and we manually evaluate them to check if they are controversial or not through multiple annotators. To classify them the content of the tweet and the received user replies were considered. A tweet is labeled controversial if the content is debatable and it expresses an idea or an opinion which generates an argument in the replies, representing opposing opinions in favor or in disagreement with the root tweet. We consider only the pages whose tweets are almost completely controversial or not controversial, and we discard all the tweets from accounts with less than 90% controversial/non-controversial tweets. The final list of the 11 controversial and 7 non-controversial selected pages are shown in Table 1. It is interesting to note that in the controversial class most of the pages are related to politics and breaking news, showing an high controversial nature of the topic, while not-controversial pages are mainly related to celebrities, and entertainment. However in our experiments since we do not use the content of the interactions, the topic of the conversation is not taken into consideration. In the subsequent analysis, we use the page as a label for the collected tweets in that page, i.e., a tweet is deemed controversial (non-controversial) if it originates from a controversial (non-controversial) classified page.

Table 1
List of Twitter pages used in our study (Dataset1).

Controversial	Non controversial
@tedcruz, @mov5stelle, @brexitwatch, @barackobama, @realdonaldtrump, @wikileaks, @berniesanders, @cnnbrk, @bbcworld, @hillaryclinton, @potus	@coldplay, @justinbieber, @cristiano, @adele, @chanel, @xbox, @nba,

Table 2
Datasets statistics.

Dataset1: *Twitter pages*			
Filtering	Root posts	Avg. users	Tot. tweets
> 2 users	1202	108	192.7*K*
> 3 users	1175 (97%)	110	192.5*K*
> 10 users	1046 (87%)	123	191.3*K*
Dataset2: *Twitter hashtags*			
Filtering	Root posts	Avg. users	Tot. tweets
> 2 users	1302	32	61.4*K*
> 3 users	1211 (93%)	34	60.5*K*
> 10 users	699 (54%)	54	54.4*K*

For each collected tweet in each page (*root post*), we reconstructed the generated discussion thread by recursively crawling the tweet's replies. The task requires a complex crawling procedure to obtain the full tree. Moreover, since we are interested in analyzing the discussion generated by each post, we restrict to the tweets that generate a conversation involving more than k users, with k=2,3 and 10. (including the author of the original post). The reply tweets are often in a different language than the language of the original tweet, including Arabic, Russian, and others. Table 2 reports the number of root posts and total reply tweets that we collect with the above procedure, with $k = 2, 3, 10$. The final dataset contains more than $190K$ tweets in total. Moreover, the table reports the average number of users who take part in the conversation for each root post. Each collected root post generates a network of replies that involves on average about 100 users.

Dataset2: ***Twitter hashtags.*** In order to be consistent with the recent literature, we also collect tweets based on controversial and non-controversial hashtags, in particular, the ones used by Garimella et al. [5]. We use four controversial (#beefban, #baltimore, #netanyahuspeech and #russia_march) and four non-controversial hashtags (#germanwings, #onedirection, #sxsw, #ultralive). For each hashtag we collect the recent posts. For each post we collect all the reply tweets and build the dataset in the same way that was described before. Statistics on this dataset are reported in Table 2. Dataset2 contains more than $60K$ tweets in total.

We note that, upon manual inspection, for many hashtags in the above-mentioned dataset, there is a mix of different behaviors depending on the context in which the hashtag is used in the tweets. Some are predominantly controversial or non-controversial, while others are mixed. Dataset2 is used as an additional test set for our model trained on Dataset1 to assess the controversial nature of popular hashtags.

4. Controversy analysis and detection

Given a social network we are interested in modeling the interactions among users and the dynamics incurring due to generated content. Users in social networks establish *friendship* or *subscription* relationships with each other, and when users interact with or publish new content their *friends* are informed. We model these relationships with a *user graph* $\mathcal{G} = (U, E)$, where U is the set of users of the network and an edge $e = (u_i, u_j) \in E$ indicates that users u_i and u_j are friends (undirected case) or that user u_i follows user u_j (directed).

Moreover, a user may publish some new content item c_i, possibly *in response to* another content item c_j authored by another user, thus generating complex threads 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 arc $r = (c_i, c_j) \in R$ 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*.

Fig. 1a shows a content *reply tree* (also referred to as just *reply tree*) present in our data, while Fig. 1b and c shows the user *reply graph* (or just *reply graph*) of two other discussion threads. Note that a social network may have several disconnected reply trees and reply graphs. Fig. 1b and c, even though are just examples, show how the network in the case of low controversy and high controversy might be really different from a structural point of view. The density of the graph in Fig. 1c for instance is higher than in Fig. 1b.

Our main hypothesis is that the structure of the user graph $\mathcal{G}$, the reply tree $\mathcal{T}$, and the reply graph $\mathcal{R}$ can be characterized by simple *motifs* of local user interactions that can be effectively exploited to distinguish between *controversial* and *non-controversial* content.

In addition to local motifs, we also explore whether baseline features (including network structure, content propagation, and temporal features) are predictors of controversy. This standard graph-based analysis is discussed in the next section while the motif-based analysis is presented in the section “Motifs.”

4.1. Baseline graph-based analysis

Structural features. The simplest structural features to extract from the user-interaction networks are the *size* in terms of *number of nodes* and *number of edges*, and the *degree distribution*.

Fig. 2a shows the distribution of the sizes of the reply tree $\mathcal{T}$ and the reply graph $\mathcal{R}$ in terms of number of nodes and number of edges for Dataset1 about Twitter pages with all the reply networks with at least 3 users involved in the conversation. To some extent, these measures are related to the popularity of the content taken into consideration. Note that in our data the sizes of $\mathcal{T}$ and $\mathcal{R}$ are very similar for both controversial and non-controversial content. This finding is in line with Smith et al. [21] that controversial content does not necessarily generate larger threads of conversation. From this, we can conclude that for distinguishing controversy *among popular topics*, just the graph sizes do not suffice.

Fig. 2b reports the average degree for the reply tree $\mathcal{T}$ and the reply graph $\mathcal{R}$. In this case, the distributions are quite different for controversial and non-controversial content. A larger average degree is observed for controversial content, suggesting that such conversations generate more engagement among users.

Propagation-based features. In order to understand how information propagates among controversial and non-controversial conversations, we investigate a number of different properties of the reply trees $\mathcal{T}$ related to information propagation. Fig. 2c shows the distribution of average and maximum cascade depths, where a cascade is defined as a path from the root to a leaf of a reply tree. The figure also shows the distribution of the maximum-size subtree among all subtrees rooted in a child of the root node. We observe that for controversial content the reply trees generally have larger depth.

Fig. 2d reports the distribution of the degree for the root, as well as the node with the larger degree excluding the root in $\mathcal{T}$. We see that in this case the controversial and non-controversial discussions have similar distributions. Nevertheless, reply trees of controversial discussions have higher probability of having a smaller root degree than non-controversial, suggesting that controversial discussions go beyond the first level of interaction.

Given the above analysis, to summarize content propagation, we decided to use the two most significant features in the content reply trees. The other features, e.g. max cascade depth, are discarded because they are strongly related to popularity. In particular:

- *average cascade depth*: the average length of root-to-leaf paths;
- *maximum relative degree*: the largest node degree excluding the root node, divided by the degree of the root.

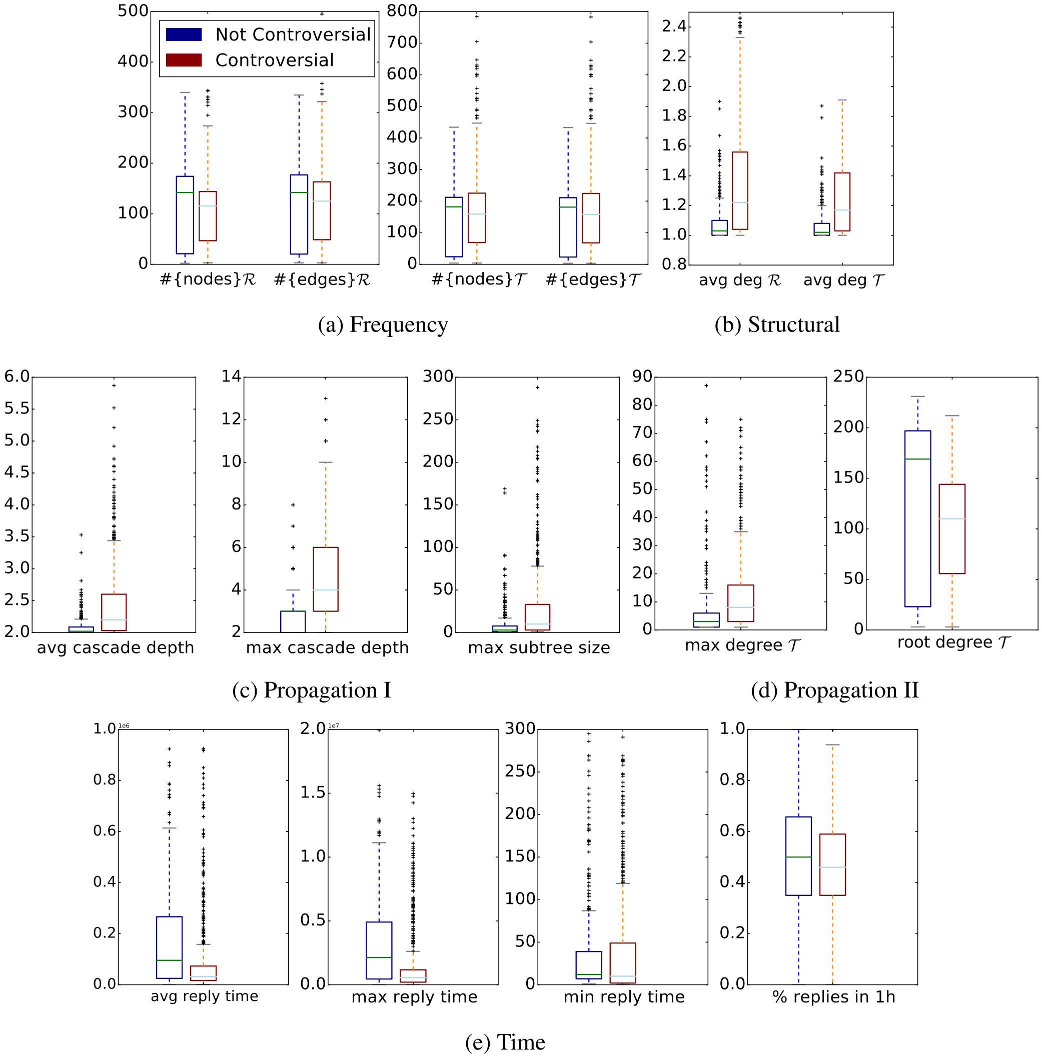
Temporal features. Considering the simple assumption that controversial topics may generate “dense” discussions in time, we analyze the time elapsed between a content item and its reply. Fig. 2e shows the distributions of minimum, maximum and average inter-reply time. Additionally, we measure the ratio of nodes in a reply tree occurring within one hour from the root. For all the measures above, there is no significant difference between controversial and non-controversial reply trees. For prediction purposes, we chose to use as features only the average inter-reply time and the ratio of replies in the first hour. Maximum and minimum inter-reply time are influenced by a single reply and for this reason they were not considered further.

4.2. Motifs

Our main hypothesis in this paper is that *local patterns* of user interaction can be used to discriminate between controversial and non-controversial discussions. This hypothesis is consistent with previous studies, where it was shown that local patterns can be used to characterize different types of networks [22,29]. As with previous work, we consider local patterns to be 2- and 3-node connected subgraphs. We refer to such patterns as *motifs*.

We consider motifs in the user graph $\mathcal{G}$ and the reply graph $\mathcal{R}$. These two graphs encompass two different kinds of information. An edge in the user graph $\mathcal{G}$ indicates that a user follows another user. These two users are likely to have similar interests and/or opinions. On the other hand, the reply graph $\mathcal{R}$ models the activity among users who may not know each other but they are willing to discuss or comment on a specific topic. In this sense, the reply graph $\mathcal{R}$ is much more dynamic and content-dependent. Antagonism between users, which can not be captured by the user graph $\mathcal{G}$ can be captured by the reply graph $\mathcal{R}$. Our basic assumption is that a combined analysis of the two graphs, $\mathcal{G}$ and $\mathcal{R}$, can lead to an improved model for controversy detection.

Dyadic motifs. We consider all possible patterns between two users in graphs $\mathcal{G}$ and $\mathcal{R}$, such that that there is at least one reply (i.e., one edge in graph $\mathcal{R}$) — otherwise the two users do not interact with each other in the discussion thread. There are seven possible configurations, which are shown in Fig. 3a. Fig. 3b shows the frequency distribution of dyadic motifs in our data. Note that patterns are mutually exclusive, therefore, pattern A where u_i replies to u_j also implies than u_j does not reply u_i and that the two users do not follow each other.



(a) Frequency

(b) Structural

(c) Propagation I

(d) Propagation II

(e) Time

Fig. 2. (a) Distribution of the number of nodes and edges in $\mathcal{T}$ and $\mathcal{R}$. (b) Distribution of average node degree in $\mathcal{T}$ and $\mathcal{R}$. (c) Distribution of avg./max. cascade depth and max. subtree size. (d) Distribution of origin degree and max. degree in $\mathcal{T}$ and $\mathcal{R}$. (e) Distribution of average, max., min. inter-reply time, and percentage of replies within one hour from the root. Non-controversial in blue (left side) vs. controversial in red (right side). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The most frequent dyadic motifs are *A* and *C*. According to Fig. 3b, it is more likely to observe a reply to a followed user in non-controversial cases. Conversely, in controversial cases it is likely to reply to a user not being followed. This confirms the intuition that controversial discussions thread interactions also among users not directly connected in the user graph $\mathcal{G}$. The features used for detecting controversial content are the frequencies of all dyadic motifs.

Triadic motifs. We also consider 3-node motifs, in particular closed triangles. As in the case of dyadic motifs, we combine structural information from the user graph $\mathcal{R}$ and the reply graph $\mathcal{G}$. Fig. 4a shows some motifs we considered. We detect a triadic motif only if there is a reply interaction among the three users. Due to the high number of possible motifs and since most motifs are relatively rare in the data, we coalesce motifs in groups. Overall, we form our set of triadic motifs by considering (*i*) the number of follow edges among the three users (Fig. 4a), (*ii*) the number of reciprocal follow edges, and (*iii*) the number of non reciprocal follow edges with opposite direction with respect to the reply edge. In total we have 20 different triadic motifs. The frequency of each motif is considered as a feature for predicting controversy.

For the lack of space we do not report the distribution for all the motifs, but generally most of the patterns we considered for closed triangles were quite rare in the dataset. Only a few of them are frequent and mostly in controversial threads, confirming the intuition that controversial discussions exhibit a more complex structure. The reason for the scarcity of complex structures is that in microblogging platforms the interactions are brief and generally involve few users. Because of the infrequency of the appearance of patterns that include more than three nodes we limited the

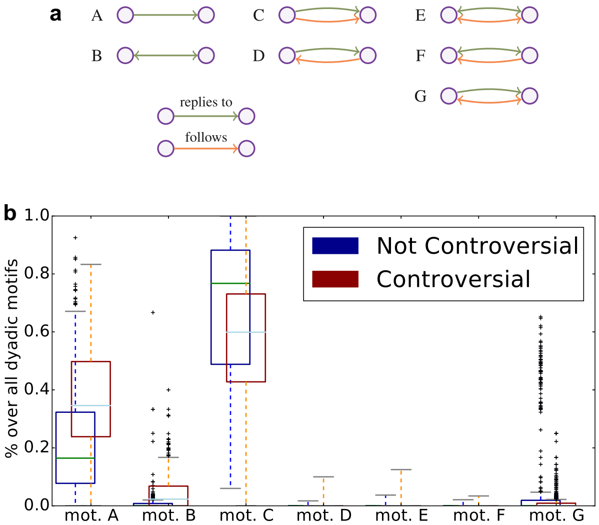


Fig. 3. (a) Dyadic motifs and (b) their frequency distribution.

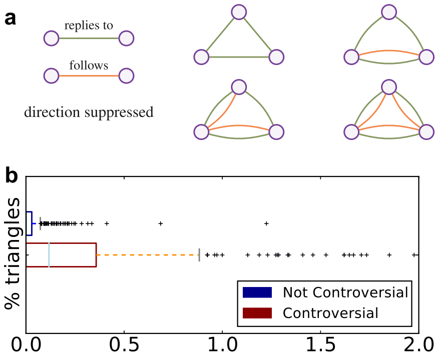


Fig. 4. (a) Triadic motifs and (b) distribution of undirected reply triangles ratio.

study to dyadic and triadic structures (the triads already showed a marginal value in our identification task). Due to this choice the network motifs used can be easily extracted. Network motifs can be easily extracted both from user interactions and from the underlying social network, and they are conceptually simple to define and very efficient to compute.

To provide additional insights on user interactions, we consider as additional feature the ratio of triangles in the reply graph $\mathcal{R}$ over the number of all possible triangles $\binom{|U|}{3}$. Again, a larger triangle ratio indicates that controversial content generates more complex discussion threads with more interactions among users and not only dyadic relations between the author of the post and the replying user, as it in the case of non-controversial situations.

We also considered “open” triadic motifs, i.e., 3-user subgraphs connected by only two replies. Such patterns did not seem to help much in predicting controversial discussions and therefore they are not considered further. The features considered in this work are shown in Table 3.

Table 3
Summary of all features.

Baseline: Structural	Avg. degree in $\mathcal{T}$ Avg. degree in $\mathcal{R}$
Baseline: Propagation	Avg. cascade depth in $\mathcal{T}$ Max. relative degree
Baseline: Temporal	Avg. inter-reply time % replies in 1h
Dyadic motifs	7 2-node motifs (shown in Fig. 3a)
Triadic motifs	20 3-node motifs Triangles ratio

5. Experiments

5.1. Detection of controversy in Twitter pages

We used the Twitter datasets presented in the data collection section. As already discussed, the Twitter pages of Dataset1 can be entirely labeled controversial or non-controversial, therefore we classify tweets according to the page it belongs. The dataset is quite balanced, with about 60% instances belonging to the controversial class and 40% to the non-controversial. Reported experiments are performed using 5-fold cross-validation and averaged over 100 trials.

We evaluated different classifiers, including AdaBoost, Logistic Regression, SVM and Random Forest, and chose AdaBoost as it resulted in the best performance. We want to detect the controversial nature of a post by analyzing user graph and reply trees. To show the relevance of detecting motifs to quantify controversy we compare the results with baseline graph-based features. We analyzed the performance by the baseline graph-based features and by using motif-based features (in addition and alone). We report the accuracy of the classifier on both controversial and non-controversial classes, and the precision, recall and F-measure with respect to the controversial class.

As shown in Table 4 the baseline approach accuracy (with structural, propagation-based and temporal features) is above 75% and increases only slightly when restricting to reply trees with

Table 4
Performance of the motif based classifier.

Filtering	Accuracy	Precision	Recall	F-measure
Baseline				
> 2 users	0.76	0.79	0.81	0.80
> 3 users	0.77	0.80	0.82	0.81
> 10 users	0.78	0.81	0.83	0.82
Baseline + dyadic motifs				
> 2 users	0.82	0.84	0.86	0.85
> 3 users	0.83	0.85	0.86	0.85
> 10 users	**0.84**	**0.86**	**0.88**	**0.87**
Baseline + dyadic and triadic motifs				
> 2 users	0.83	0.85	0.86	0.85
> 3 users	0.84	0.86	0.85	0.86
> 10 users	**0.85**	**0.87**	**0.88**	**0.87**
Dyadic motifs only				
> 2 users	0.75	0.77	0.82	0.80
> 3 users	0.75	0.77	0.82	0.80
> 10 users	0.77	0.79	0.84	0.82
Triadic motifs only				
> 2 users	0.73	0.89	0.62	0.73
> 3 users	0.74	0.88	0.64	0.74
> 10 users	0.77	0.89	0.71	0.79
Dyadic + Triadic motifs only				
> 2 users	0.77	0.81	0.80	0.81
> 3 users	0.77	0.81	0.80	0.80
> 10 users	0.80	0.84	0.82	0.83

Table 5
Feature importance (filtering > 10 users).

Feature	Error reduction
(1) Avg. inter-reply time	0.18
(2) Max. relative degree	0.16
(3) Motif *A*	0.14
(4) % Replies within 1 h	0.08
(5) Motif *B*	0.08
(6) Motif *G*	0.06
(7) Triangles ratio	0.04
(8) Triadic motif	0.04

more than 10 users. With the addition of dyadic motifs, all the performance figures are significantly improved. Note that the precision of the algorithm improves in both controversial and non-controversial classes. The addition of triadic motifs leads to the best results, but the improvement is only marginal. This is because, as discussed in the previous section, triads are infrequent: even if conveying relevant information, they may help in improving the classification of a limited number of instances. The best results highlighted in boldface in Table 4 are statistically significant (t-test with p-values $\ll 0.01$) w.r.t. baseline features. Using dyadic motifs alone, moreover, the accuracy of the model is comparable with the baseline, with a limited improvement if we add the triadic patterns.

In Table 5 we report the 8 most relevant features exploited by the AdaBoost model according to their contribution in the error reduction. Temporal features are important to detect controversy. The first feature is the average inter-reply time, and the fourth is the ratio of replies posted within one hour of the original tweet: when the discussion is polarized people tend to reply in a shorter time. This result is in line with other contexts. For example, it is known that temporal features play the main role to predict popularity [30]. The second most important feature is the maximum relative degree, i.e., the maximum degree normalized by the root node degree. In non-controversial reply trees, the root is the only node with a large degree, i.e., the node attracting most of the reply activity.

The other features among the top-6 are dyadic motifs. The most relevant being motif *A*, which corresponds to a user u_i replying to u_j without any following relationship among the two. We deduce that controversial threads create engagement among users not being directly connected in the social network. On the other hand, the fact that motif *C* is not relevant (where a user replies to a follower), suggests that it is less likely to have controversial discussions among friends. Interestingly, dyadic patterns seem to be more relevant than propagation-based features. For instance, the depth of the cascades, which was expected to model the complexity of the interactions, is not among the top-8 features. Presumably, complex propagation features are superseded by the simple motif patterns.

Finally, the last two important features are based on triangles. In particular the relevance of the triangle-ratio feature suggests that triadic patterns are able to grasp interactions occurring in controversial discussions. However it is harder to draw any conclusion on the role of specific triads patterns, due to their low frequency. The most significant specific triadic pattern included in the list in Table 5 is a close reply triangle with two follow edges: one reciprocal and one not reciprocal with the same direction of the underlying reply edge. Since triadic patterns provide a limited contribution to the classifier, we conclude that dyadic motifs are already effective, and there is not much information that can be extracted based on specific triadic motifs.

5.2. Dynamic tracking of controversy

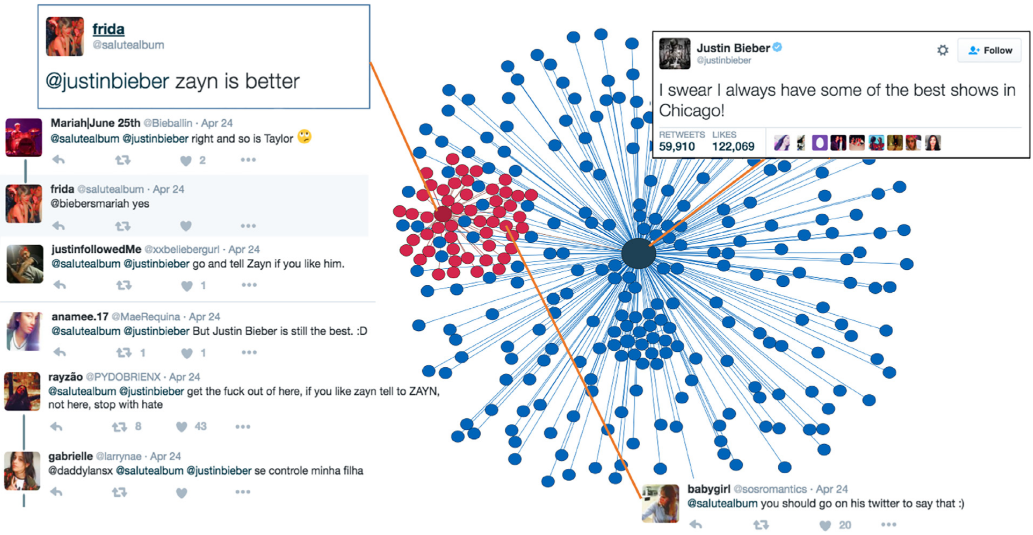
We found it is not always appropriate to classify a reply tree as controversial or not. This is because each reply may generate unexpected reaction. For instance, there may be sub-threads of controversy, within a non-controversial discussion. To test this intuition, we analyzed the direct replies of the *origin* tweets that were classified as non-controversial. This can be achieved easily as the proposed approach can be applied to any tweet given its reply tree, or in this case, its reply sub-tree. By applying the model discussed in the previous section, we found that about 7% of the direct-reply sub-trees of a non-controversial tweet are controversial.

One such example is shown in Fig. 5, illustrating the reply tree of a post by Justin Bieber. A majority of the replies are not controversial and are written by his fans with compliments and expressions of affection and love. However, the proposed algorithm detected as controversial one sub-tree (highlighted in red) generated by a reply in support to another singer: “Zayn is better.” This post generated a subtree with animated discussion among fans. A similar case was found for Cristiano Ronaldo’s profile, where a number of users started discussion about his rivalry with Messi.

Both of the previous examples are typical cases in which the controversial portion of the discussion is limited to a few branches, and its detection might be challenging. We claim that the proposed approach, based on local motifs can successfully detect small controversial sub-threads.

5.3. Hashtags evaluation

Since on Twitter, topics are often identified through hashtags, we tested the proposed method on tweets mentioning a given hashtag (Dataset2), obtained from the previous work [5]. Table 6 shows the fraction of controversial posts per hashtag, as detected by our model. The smallest fraction of controversial discussions is found with #sxsw and #ultralive hashtags (related to music events), where most conversations are expected to happen among supporters of the same music band. The most controversial discussions are found with the #beefban, #onedirection, #netanyahu,



frida @salutealbum
@justinbieber zayn is better

Mariah|June 25th @Bieballin · Apr 24
@salutealbum @justinbieber right and so is Taylor 🙄
2

frida @salutealbum · Apr 24
@biebersmariah yes

justinfollowedMe @xxbeliebergurl · Apr 24
@salutealbum @justinbieber go and tell Zayn if you like him.
1

anamee.17 @MaeRequina · Apr 24
@salutealbum @justinbieber But Justin Bieber is still the best. :D
1 1

rayzão @PYDOBRIENX · Apr 24
@salutealbum @justinbieber get the fuck out of here, if you like zayn tell to ZAYN, not here, stop with hate
8 43

gabrielle @larrynae · Apr 24
@daddylansx @salutealbum @justinbieber se controle minha filha

Justin Bieber @justinbieber
Follow
I swear I always have some of the best shows in Chicago!
RETWEETS 59,910 LIKES 122,069

babygirl @sosromantics · Apr 24
@salutealbum you should go on his twitter to say that :)
20

Fig. 5. A controversial reply sub-tree (in red) originated by a non-controversial post (in blue) by Justin Bieber. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

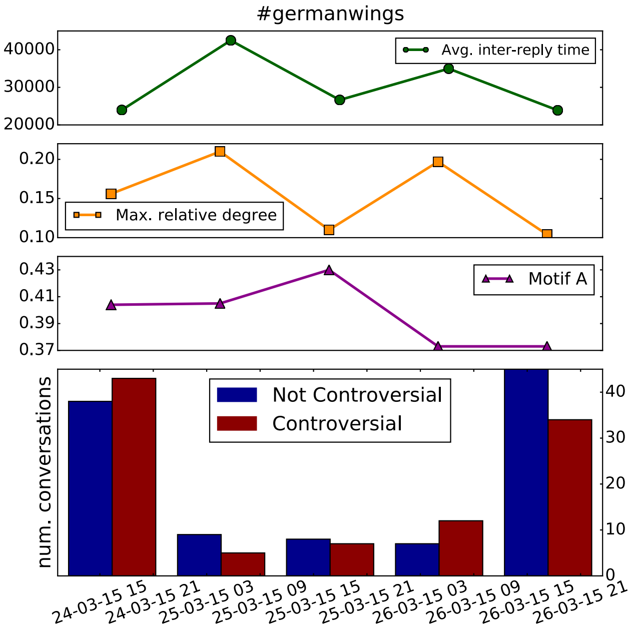


Fig. 6. Distribution of controversial (red) vs. non-controversial (blue) posts and top-3 features values over time for the #germanwings hashtag. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

#baltimore hashtags. The classification of these hashtags as controversial is in line with the previous results [5], with the exception of #onedirection for which we detected antagonist replies, upon manual inspection. Most of the hashtags exhibit a mixed behavior as far as controversy is concerned.[1] Indeed, simply counting the number of tweets classified as controversial is a quite naïve approach, strongly dependent on different factors, such as the daily volume of tweets, on external events, and many others. For these reasons, we believe that it is more interesting to study how the controversy related to a given hashtag evolves over time.

Fig. 6 shows the evolution of the controversy for the #germanwings hashtag. Note that some hours after the accident happened on March 24 the majority of threads are controversial. In the evening the discussions become less controversial and mainly about sorrow and condolences. An interesting increase of the

[1] E.g.: A controversial tweet id=580330769912061953 and a non-controversial tweet id=580361863449444352 about #germanwings.

Table 6
Hashtag controversy classification.

Hashtag	Ratio of controversial posts
sxsw	0.32
Germanwings	0.49
Beefban	0.70
Netanyahu	0.55
Ultralive	0.29
Onedirection	0.61
Baltimore	0.58
Russia-march	0.46

controversy level is registered the next day, until details about the accident were released. Then the discussion becomes predominantly non-controversial showing that the audience has digested the news. We highlight that the level of controversy is anticorrelated with the frequency for motif *A*, thus confirming the prediction power of the proposed motifs. Moreover, we showed in the figure also the trend for other significant features used by the classification model. The average inter-reply time and the maximum relative degree are trends are very similar but the correlation with the classification results is not easily evident, thus explaining the importance of combining many different features to get a useful classification.

6. Conclusion

We proposed a novel approach based on local graph motifs for identifying controversy on online social networks. The proposed method is language independent and exploits local patterns of user interactions to detect controversial threads of discussion. Given a content item, users reply to each other generating different configurations of the reply graph. We investigated local motifs extracted from this graph and from the user friendship graph. Such motifs correspond to different interaction patterns among two users, which may be linked by a possibly reciprocal reply action and by a possibly reciprocal friendship relationship. Similar motifs regarding the interaction of three users were considered.

We proved on a benchmark Twitter dataset that such motifs are more powerful in predicting controversy than other baseline frequently used graph properties such as cascade depth. Specifically dyadic patterns seem to be more relevant than structural features to detect controversy. We observed that in most cases controversy arise when users participate to discussions beyond their social circles. This means that it is less likely to have controversial discussions among friends. Finally, as the proposed motifs can be easily extracted from any reply tree or sub-tree, we experimented with the use of such patterns in monitoring the evolution of discussions and sub-discussions over time. Indeed, we found that a topic of discussion develops over time changing its level of controversy depending on different sub-topics or on external events (e.g., news). About 7% of the direct-reply sub-trees of a non-controversial tweet are detected as controversial.

Therefore, a fine-grained analysis, as provided by the proposed local motifs, is necessary for a better understanding of controversy in online social networks.

Acknowledgments

This work was partially supported by the EC H2020 Program INFRAIA-1-2014-2015 *SoBigData: Social Mining & Big Data Ecosystem* (654024).

References

[1] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3) (2010) 75–174.
[2] E. Bakshy, I. Rosenn, C. Marlow, L. Adamic, The role of social networks in information diffusion, in: Proceedings of the WWW, ACM, 2012, pp. 519–528.
[3] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a social network or a news media? in: Proceedings of the WWW, ACM, 2010, pp. 591–600.
[4] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (7) (2007) 1019–1031.
[5] K. Garimella, G. De Francisci Morales, A. Gionis, M. Mathioudakis, Quantifying controversy in social media, in: Proceedings of the WSDM, ACM, 2016, pp. 33–42.
[6] H. Lu, J. Caverlee, W. Niu, Biaswatch: a lightweight system for discovering and tracking topic-sensitive opinion bias in social media, in: Proceedings of the CIKM, ACM, 2015, pp. 213–222.
[7] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, A. Flammini, Political Polarization on Twitter, in: Proceedings of the ICWSM, 2011.
[8] Y. Mejova, A.X. Zhang, N. Diakopoulos, C. Castillo, Controversy and sentiment in online news, Symp. Comput. Journal. (2014).
[9] M. Coletto, K. Garimella, A. Gionis, C. Lucchese, A motif-based approach for identifying controversy, in: Proceedings of the ICWSM, 2017.
[10] S. Dori-Hacohen, J. Allan, Detecting controversy on the web, in: Proceedings of the CIKM, ACM, 2013, pp. 1845–1848.
[11] L.A. Adamic, N. Glance, The political blogosphere and the 2004 us election: divided they blog, in: Proceedings of the LinkKDD, 2005, pp. 36–43.
[12] Y. Choi, Y. Jung, S.-H. Myaeng, Identifying controversial issues and their sub–topics in news articles, in: Intelligence and Security Informatics, Springer, 2010, pp. 140–153.
[13] J. An, D. Quercia, J. Crowcroft, Partisan sharing: Facebook evidence and societal consequences, in: COSN, 2014, pp. 13–24.
[14] P.H.C. Guerra, W. Meira Jr, C. Cardie, R. Kleinberg, A measure of polarization on social media networks based on community boundaries, in: Proceedings of the ICWSM, 2013.
[15] A. Morales, J. Borondo, J. Losada, R. Benito, Measuring political polarization: Twitter shows the two sides of Venezuela, Chaos 25 (3) (2015).
[16] M. Coletto, C. Lucchese, S. Orlando, R. Perego, Polarized user and topic tracking in twitter, in: Proceedings of the SIGIR, Pisa, Italy, 2016.
[17] P. Cogan, M. Andrews, M. Bradonjic, W.S. Kennedy, A. Sala, G. Tucci, Reconstruction and analysis of Twitter conversation graphs, in: Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research, ACM, 2012, pp. 25–31.
[18] R. Nishi, T. Takaguchi, K. Oka, T. Maehara, M. Toyoda, K.-i. Kawarabayashi, N. Masuda, Reply trees in twitter: data analysis and branching process models, Soc. Netw. Anal. Min. 6 (1) (2016) 1–13.
[19] M. De Domenico, A. Lima, P. Mougel, M. Musolesi, The anatomy of a scientific rumor, Sci. Rep. 3 (2013) 2980.
[20] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the WWW, ACM, 2011, pp. 675–684.
[21] L.M. Smith, L. Zhu, K. Lerman, Z. Kozareva, The role of social media in the discussion of controversial topics, in: Proceedings of the SocialCom, IEEE, 2013, pp. 236–243.
[22] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, Science 298 (5594) (2002) 824–827.
[23] A.R. Benson, D.F. Gleich, J. Leskovec, Higher-order organization of complex networks, Science 353 (6295) (2016) 163–166.
[24] K. Allen, G. Carenini, R.T. Ng, Detecting disagreement in conversations using pseudo-monologic rhetorical structure., in: Proceedings of the EMNLP, 2014, pp. 1169–1180.
[25] Z. Chen, J. Berger, When, why, and how controversy causes conversation, J. Consum. Res. 40 (3) (2013) 580–593.
[26] A. Kanavos, I. Perikos, P. Vikatos, I. Hatzilygeroudis, C. Makris, A. Tsakalidis, Conversation emotional modeling in social networks, in: Proceedings of the IEEE 26th International Conference on Tools with Artificial Intelligence, IEEE, 2014, pp. 478–484.
[27] M.B. Zafar, K.P. Gummadi, C. Danescu-Niculescu-Mizil, Message impartiality in social media discussions, in: Proceedings of the ICWSM, 2016.
[28] L. Barbosa, J. Feng, Robust sentiment detection on twitter from biased and noisy data, in: ICCL: Posters, Association for Computational Linguistics, 2010, pp. 36–44.
[29] I. Bordino, D. Donato, A. Gionis, S. Leonardi, Mining large networks with subgraph counting, in: Proceedings of the Eighth IEEE ICDM, IEEE, 2008, pp. 737–742.
[30] B. Shulman, A. Sharma, D. Cosley, Predictability of popularity: Gaps between prediction and understanding, in: Proceedings of the ICWSM, 2016.

Mauro Coletto is a research fellow at DAIS, Ca' Foscari University, and he received his Ph.D. degree from IMT - Institute for Advanced Studies (Lucca) in the track "Computer, Decision, and Systems Science". He graduated in Information Management Engineering at the University of Udine in 2012. During his doctoral studies at IMT, he has worked in collaboration with CNR (ISTI-HPC) on the following research topics: Web Mining, Online Social Networks, and Social Media Analysis.

Kiran Garimella is a Ph.D. student at Aalto University. His research focuses on identifying and combating filter bubbles on social media. Previously he worked as a Research Engineer at Yahoo Research, QCRI and as an Research Intern at LinkedIn and Amazon. His research on polarization on social media received the best student paper awards at WSDM 2017 and Webscience 2017.

Aristides Gionis is Professor in the Department of Computer Science of Aalto University, leading the Data Mining Group. His research focuses on data mining and algorithmic data analysis. He is particular interested in algorithms for graphs, social-network analysis, and algorithms for web-scale data. He was a senior research scientist in Yahoo! Research, and previously an Academy of Finland postdoctoral scientist in the University of Helsinki. He obtained his Ph.D. from Stanford University in 2003. He had the pleasure to work as a summer intern in Microsoft Research, AT&T Labs, and Bell Labs Lucent Technologies.

Claudio Lucchese is Associate Professor with the Dipartimento di Scienze Ambientali, Informatica e Statistica at the Universit Ca' Foscari di Venezia. Between 2007 and 2017 he was researcher with the I.S.T.I. "A. Faedo" C.N.R.. Prior to joining C.N.R., he received his MS.c. and Ph.D. in Computer Science from the Universit Ca' Foscari di Venezia in 2003 and 2008, respectively. His main research activities are in the areas of data mining techniques for information retrieval and large-scale data processing.

Publication 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.

Quantifying Controversy on Social Media

KIRAN GARIMELLA, Aalto University
GIANMARCO DE FRANCISCI MORALES, Qatar Computing Research Institute
ARISTIDES GIONIS, Aalto University
MICHAEL MATHIOUDAKIS, Aalto University

CCS Concepts: • **Information systems** → **Social networking sites**; *Social networks*; • **Human-centered computing** → *Social networks*; • **Applied computing** → Sociology;

Additional Key Words and Phrases: controversy, polarization, echo chambers, filter bubble, twitter

ACM Reference format:
Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. Quantifying Controversy on Social Media. 1, 1, Article 1 (September 2017), 26 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Given their widespread diffusion, online social media are becoming increasingly important in the study of social phenomena such as peer influence, framing, bias, and controversy. Ultimately, we would like to understand how users perceive the world through the lens of their social media feed. However, before addressing these advanced application scenarios, we first need to focus on the fundamental yet challenging task of distinguishing whether a topic of discussion is controversial. Our work is motivated by interest in observing controversies at societal level, monitoring their evolution, and possibly understanding which issues become controversial and why.

The study of controversy in social media is not new; there are many previous studies aimed at identifying and characterizing controversial issues, mostly around political debates [1, 10, 38, 39] but also for other topics [26]. And while most recent papers have focused on Twitter [10, 26, 38, 39], controversy in other social-media platforms, such as blogs [1] and opinion fora [2], has also been analyzed.

However, most previous papers have severe limitations. First, the majority of previous studies focus on controversy regarding political issues, and, in particular, are centered around long-lasting major events, such as elections [1, 10]. More crucially, most previous works can be characterized as *case studies*, where controversy is identified in a single carefully-curated dataset, collected using ample domain knowledge and auxiliary domain-specific sources (e.g., an extensive list of hashtags regarding a major political event, or a list of left-leaning and right-leaning blogs).

We aim to overcome these limitations. We develop a framework to identify controversy regarding topics in any domain (e.g., political, economical, or cultural), and without prior domain-specific knowledge about the topics in question. Within the framework, we quantify the controversy associated with each topic, and thus compare different topics in order to find the most controversial

DOI: 10.1145/nnnnnnn.nnnnnnn

ones. Having a framework with these properties allows us to deploy a system in-the-wild, and is valuable for building real-world applications.

In order to enable such a versatile framework, we work with topics that are defined in a light-weight and domain-agnostic manner. Specifically, when focusing on Twitter, a topic is specified as a text query. For example, "#beefban" is a special keyword (a "hashtag") that Twitter users employed in March 2015 to signal that their posts referred to a decision by the Indian government about the consumption of beef meat in India. In this case, the query "#beefban" defines a topic of discussion, and the related activity consists of all posts that contain the query, or other closely related terms and hashtags, as explained in Section 4.1.

We represent a topic of discussion with a *conversation graph*. In such a graph, vertices represent users, and edges represent conversation activity and interactions, such as *posts*, *comments*, *mentions*, or *endorsements*. Our working hypothesis is that it is possible to analyze the conversation graph of a topic to reveal how controversial the topic is. In particular, we expect the conversation graph of a controversial topic to have a *clustered structure*. This hypothesis is based on the fact that a controversial topic entails different sides with opposing points of view, and individuals on the same side tend to endorse and amplify each other's arguments [1, 2, 10].

Our main contribution is to test this hypothesis. We achieve this by studying a large number of candidate features, based on the following *aspects* of activity: (*i*) *structure of endorsements*, i.e., who agrees with whom on the topic, (*ii*) *structure of the social network*, i.e., who is connected with whom among the participants in the conversation, (*iii*) *content*, i.e., the keywords used in the topic, (*iv*) *sentiment*, i.e., the tone (positive or negative) used to discuss the topic. Our study shows that, except from content-based features, all the other ones are useful in detecting controversial topics, to different extents. Particularly for Twitter, we find the endorsement features (i.e., retweets) to be the most useful.

The extracted features are then used to compute the *controversy score* of a topic. We offer a systematic definition and provide a thorough evaluation of measures to quantify controversy. We employ a broad range of topics, both controversial and non-controversial ones, on which we evaluate several measures, either defined in this paper or coming from the literature [26, 39]. We find that one of our newly-proposed measure, based on *random walks*, is able to discriminate controversial topics with great accuracy. In addition, it also generalizes well as it agrees with previously-defined measures when tested on datasets from existing work. We also find that the *variance* of the sentiment expressed on a topic is a reliable indication of controversy.

The approach to quantifying controversy presented in this paper can be condensed into a three-stage pipeline: (*i*) building a *conversation graph* among the users who contribute to a topic, where edges signify that two users are in agreement, (*ii*) identifying the potential sides of the controversy from the graph structure or the textual content, and (*iii*) quantifying the amount of controversy in the graph.

The rest of this paper is organized as follows. Section 2 discusses how this work fills gaps in the existing literature. Subsequently, Section 3 provides a high level description of the pipeline for quantifying controversy of a topic, while Sections 4, 5, and 6 detail each stage. Section 7 shows how to extend the controversy measures from topics to users who participate in the discussion. We report the results of an extensive empirical evaluation of the proposed measures of controversy in Section 8. Section 9 extends the evaluation to a few measures that do not fit the pipeline. We conclude in Section 10 with a discussion on possible improvements and directions for future work, as well as lessons learned from carrying out this study.

2 RELATED WORK

Analysis of controversy 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 and Glance [1] study the link patterns and discussion topics of political bloggers, focusing on blog posts on the U.S. presidential election of 2004. 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. [10], who also study controversy in political communication regarding congressional midterm elections. Using data from Twitter, Conover et al. [10] 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. In another recent work related to controversy analysis in political discussion, Mejova et al. [38] identify a significant correlation between controversial issues and the use of negative affect and biased language.

The papers mentioned so far study controversy in the political domain, and provide case studies centered around long-lasting major events, such as presidential elections. In this paper, we aim to identify and quantify controversy for any topic discussed in social media, including short-lived and ad-hoc ones (for example, see topics in Table 2). The problem we study has been considered by previous work, but the methods proposed so far are, to a large degree, domain-specific.

The work of Conover et al. [10], discussed above, 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. [26] 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. Differently from these studies, our contribution consists in providing an extensive study of a large number of measures, including the ones proposed earlier, and demonstrating clear improvement over those. We also aim at quantifying controversy in diverse and in-the-wild settings, rather than carefully-curated domain-specific datasets.

In a recent study, Morales et al. [39] quantify polarity via the propagation of opinions of influential users on Twitter. They validate their measure with a case study from Venezuelan politics. Again, our methods are not only more general and domain agnostic, but they provide more intuitive results. In a different approach, Akoglu [2] proposes a polarization metric that uses signed bipartite opinion graphs. The approach differs from ours as it relies on the availability of this particular type of data, which is not as readily available as social-interaction graphs.

Similarly to the papers discussed above, in our work we quantify controversy based on the graph structure of social interactions. In particular, we assume that controversial and polarized topics induce 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 quantified in many recent studies [4, 18, 25].

A different direction for quantifying controversy followed by Choi et al. [8] and Mejova et al. [38] 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 controversy in a domain-agnostic manner, text analysis has its limitations. Nevertheless, we experiment with incorporating content features in our approach.

A summary of related work along different dimensions is summarized in Table 1. As we mention above, most existing work to date tries to $identify$ controversial topics as case studies on a particular topic, either using content or networks of interactions. Our work is one of the few that $quantifies$

Table 1. Summary of related work for identfying/quantifying controversial topics

Paper	Identifying	Quantifying	Content	Network
Choi et al. [8]	✓		✓	
Popescu et al. [44]	✓		✓	
Mejova et al. [38]	✓		✓	
Klenner et al. [32]	✓		✓	
Tsytsarau et al. [48]	✓		✓	
Dori et al. [14]	✓		✓	
Jang et al. [28]	✓		✓	
Conover et al. [10]	✓			✓
Coletto et al. [9]	✓			✓
Akoglu et al. [2]	✓			✓
Amin et al. [3]	✓			✓
Guerra et al. [26]	✓	✓		✓
Morales et al. [39]	✓	✓		✓
Garimella et al. [20]	✓	✓		✓

the degree of controversy using language and domain independent methods. We show in Section 8 that our method outperforms [26, 39].

Finally, our findings on controversy have many potential applications on news-reading and public-debate scenarios. For instance, quantifying controversy can provide a basis for analyzing the "news diet" of readers [33, 34], offering the chance of better information by providing recommendations of contrarian views [40], deliberating debates [16], and connecting people with opposing opinions [15, 24].

3 PIPELINE

Our approach to measuring controversy is based on a systematic way of characterizing social media activity. We employ a pipeline with three stages, namely *graph building*, *graph partitioning*, and *measuring controversy*. The final output of the pipeline is a value that measures how controversial a topic is, with higher values corresponding to higher degree of controversy. We provide a high-level description of each stage here and more details in the sections that follow.

3.1 Building the Graph

The purpose of this stage is to build a *conversation graph* that represents activity related to a single *topic* of discussion. In our pipeline, a topic is operationalized as a set of related hashtags (details in §4.1), and the social media activity related to the topic consists of those items (e.g., posts) that

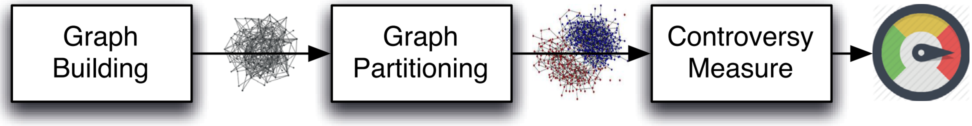


Fig. 1. Block diagram of the pipeline for computing controversy scores.

match this set of hashtags. For example, in the context of Twitter, the query might consist simply of a keyword, such as “#ukraine", in which case the related activity consists of all tweets that contain that keyword, or related tags such as #kyiv and #stoprussianaggression. Even though we describe textual queries in standard document-retrieval form, in principle queries can take other forms, as long as they are able to induce a graph from the social media activity (e.g., RDF queries, or topic models).

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, an edge between two vertices represents *endorsment*, *agreement*, or *shared point of view* between the corresponding users. Section 4 details several ways to build such a graph.

3.2 Partitioning the Graph

In the second stage, the resulting conversation graph is fed into a *graph partitioning* algorithm to extract *two* partitions (we defer considering multi-sided controversies to a further study). 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?” Section 5 describes this stage in further detail. If indeed there are two sides which do not agree with each other –a controversy– then the two partitions should be loosely connected to each other, given the semantic of the edges. This property is captured by a measure computed in the third and final stage of the pipeline.

3.3 Measuring Controversy

The third and last stage takes as input the graph built by the first stage and partitioned by the second stage, and computes the value of a *controversy measure* that characterizes how controversial the topic is. Intuitively, a controversy measure aims to capture how separated the two partitions are. We test several such measures, including ones based on random walks, betweenness centrality, and low-dimensional embeddings. Details are provided in Section 6.

4 GRAPH BUILDING

This section provides details about the different approaches we follow to build graphs from raw data. We use posts on Twitter to create our datasets.[1] 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. Following the procedure described in Section 3.1, we specify a set of queries (indicating topics), and build one graph for each query. We choose a set of topics balanced between controversial and non-controversial ones, so as to test for both false positives and false negatives.

We use Twitter hashtags as *queries*. Users commonly employ hashtags to indicate the topic of discussion their posts pertain to. Then, we define a *topic* as the set of hashtags related to the given query. Among the large number of hashtags that appear in the Twitter stream, we consider those that were trending during the period from Feb 27 to Jun 15, 2015. By manual inspection we find that most trending hashtags are not related to controversial discussions [19].

We first manually pick a set of 10 hashtags that we know represent *controversial* topics of discussion. All hashtags in this set have been widely covered by mainstream media, and have generated ample discussion, both online and offline. Moreover, to have a dataset that is balanced between controversial and non-controversial topics, we sample another set of 10 hashtags that

[1] From the full Twitter firehose stream.

Table 2. Datasets statistics: hashtag, sizes of the follow and retweet graphs, and description of the event. The top group represent controversial topics, while the bottom one represent non-controversial ones.

Hashtag	# Tweets	Retweet graph		Follow graph		Description and collection period (2015)
		$\lvert V\rvert$	$\lvert E\rvert$	$\lvert V\rvert$	$\lvert E\rvert$	
#beefban	422 908	21 590	30 180	9525	204 332	Government of India bans beef, Mar 2–5
#nemtsov	371 732	43 114	77 330	17 717	155 904	Death of Boris Nemtsov, Feb 28–Mar 2
#netanyahuspeech	1 196 215	122 884	280 375	49 081	2 009 277	Netanyahu's speech at U.S. Congress, Mar 3–5
#russia_march	317 885	10 883	17 662	4844	42 553	Protests after death of Boris Nemtsov ("march"), Mar 1–2
#indiasdaughter	776 109	68 608	144 935	38 302	131 566	Controversial Indian documentary, Mar 1–5
#baltimoreriots	1 989 360	289 483	432 621	214 552	690 944	Riots in Baltimore after police kills a black man, Apr 28–30
#indiana	972 585	43 252	74 214	21 909	880 814	Indiana pizzeria refuses to cater gay wedding, Apr 2–5
#ukraine	514 074	50 191	91 764	31 225	286 603	Ukraine conflict, Feb 27–Mar 2
#gunsense	1 022 541	30 096	58 514	17 335	841 466	Gun violence in U.S., Jun 1–30
#leadersdebate	2 099 478	54 102	136 290	22 498	1 211 956	Debate during the U.K. national elections, May 3
#sxsw	343 652	9304	11 003	4558	91 356	SXSW conference, Mar 13–22
#1dfamheretostay	501 960	15 292	26 819	3151	20 275	Last OneDirection concert, Mar 27–29
#germanwings	907 510	29 763	39 075	2111	7329	Germanwings flight crash, Mar 24–26
#mothersday	1 798 018	155 599	176 915	2225	14 160	Mother's day, May 8
#nepal	1 297 995	40 579	57 544	4242	42 833	Nepal earthquake, Apr 26–29
#ultralive	364 236	9261	15 544	2113	16 070	Ultra Music Festival, Mar 18–20
#FF	408 326	5401	7646	3899	63 672	Follow Friday, Jun 19
#jurassicworld	724 782	26 407	32 515	4395	31 802	Jurassic World movie, Jun 12-15
#wcw	156 243	10 674	11 809	3264	23 414	Women crush Wednesdays, Jun 17
#nationalkissingday	165 172	4638	4816	790	5927	National kissing day, Jun 19

represent *non-controversial* topics of discussion. These hashtags are related mostly to "soft news" and entertainment, but also to events that, while being impactful and dramatic, did not generate large controversies (e.g., #nepal and #germanwings). In addition to our intuition that these topics are non-controversial, we manually check a sample of tweets, and we are unable to identify any clear instance of controversy.[2]

As a first step, we now describe the process of expanding a single hashtag into a set of related hashtags which define the topic. The goal of this process is to broaden the definition of a topic, and ultimately improve the coverage of the topic itself.

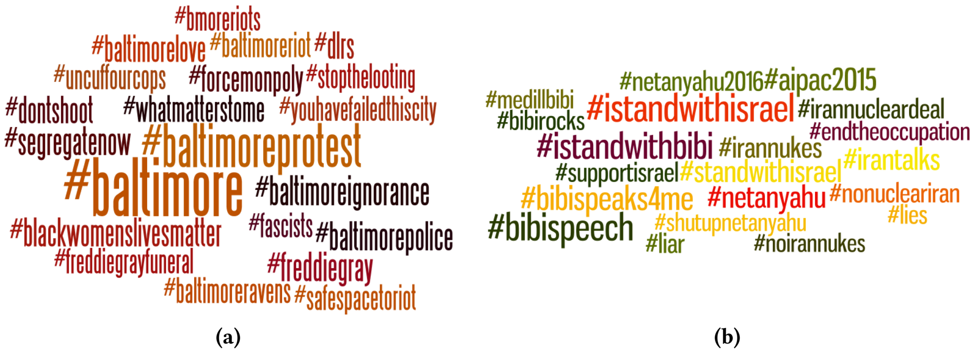
4.1 From hashtags to topics

In the literature, a topic is often defined by a single hashtag. However, this choice might be too restrictive in many cases. For instance, the opposing sides of a controversy might use different hashtags, as the hashtag itself is loaded with meaning and used as a means to express their opinion. Using a single hashtag may thus miss part of the relevant posts.

To address this limitation, we extend the definition of topic to be more encompassing. Given a *seed* hashtag, we define a topic as a set of related hashtags, which co-occur with the seed hashtag. To find related hashtags, we employ (and improve upon) a recent clustering algorithm tailored for the purpose [17].

Feng et al. [17] develop a simple measure to compute the similarity between two hashtags, which relies on co-occurring words and hashtags. The authors then use this similarity measure to find closely related hashtags and define clusters. However, this simple approach presents one drawback, in that very popular hashtags such as #ff or #follow co-occur with a large number of hashtags. Hence, directly applying the original approach results in extremely noisy clusters. Since the quality

[2] Code and networks used in this work are available at http://github.com/gvrkiran/controversy-detection.



(a) **(b)**

Fig. 2. Sets of related hashtags for the topics (a) #baltimoreriots and (b) #netanyahuspeech.

of the topic affects critically the entire pipeline, we want to avert this issue and ensure minimal noise is introduced in the expanded set of hashtags.

Therefore, we improve the basic approach by taking into account and normalizing for the popularity of the hashtags. Specifically, we compute the document frequency of all hashtags on a random 1% sample of the Twitter stream[3], and normalize the original similarity score between two hashtags by the inverse document frequency. The similarity score is formally defined as

$$sim(h_s, h_t) = \frac{1}{1 + \log(df(h_t))} \left(\alpha \ \cos(W_s, W_t) + (1 - \alpha) \ \cos(H_s, H_t)\right), \tag{1}$$

where h_s is the seed tag, h_t is the candidate tag, W_x and H_x are the sets of words and hashtags that co-occur with hashtag h_x, respectively, cos is the cosine similarity between two vectors, df is the document frequency of a tag, and α is a parameter that balances the importance of words compared to hashtags in a post.

By using the similarity function in Equation 1, we retrieve the top-k most similar hashtags to a given seed. The set of these hashtags along with the initial seed defines the topic for the given seed hashtag. The topic is used as a filter to get all tweets which contain at least one of the hashtags in the topic. In our experiments we use $\alpha = 0.3$ (as proposed by Feng et al. [17]) and $k = 20$.

Figure 2 shows the top-20 most similar hashtags for two different seeds: (a) #baltimoreriots, that identifies the discussion around the Baltimore riots against police violence in April 2015 and (b) #netanyahuspeech, that identifies the discussion around Netanyahu's speech at the US congress in March 2015. By inspecting the sets of hashtags, it is possible to infer the nature of the controversy for the given topic, as both sides are represented. For instance, the hashtags #istandwithisrael and #shutupbibi represent opposing sides in the dicussion raised by Netanyahu's speech. Both hashtags are recovered by our approach when #netanyahuspeech is provided as the seed hashtag. It is also clear why using a single hashtag is not sufficient to define a topic: the same user is not likely to use both #safespacetoriot and #segregatenow, even though the two hashtags refer to the same event (#baltimoreriots).

4.2 Data aspects

For each topic, we retrieve all tweets that contain one of its hashtags and that are generated during the observation window. We also ensure that the selected hashtags are associated with a

[3] from the Twitter Streaming API https://dev.twitter.com/streaming/reference/get/statuses/sample

large enough volume of activity. Table 2 presents the final set of seed hashtags, along with their description and the number of related tweets.[4] For each topic, we build a graph G where we assign a vertex to each user who contributes to it, and generate edges according to one of the following four approaches, which capture different *aspects* of the data source.

1. Retweet graph. Retweets typically indicate endorsement.[5] Users who retweet signal endorsement of the opinion expressed in the original tweet by propagating it further. Retweets are not constrained to occur only between users who are connected in Twitter's social network, but users are allowed to retweet posts generated by any other user.

We select the edges for graph G based on the retweet activity in the topic: an edge exists between two users u and v if there are at least *two* ($\tau = 2$) retweets between them that use the hashtag, irrespective of direction. We remark that, in preliminary experimentation with this approach, building the retweet graph with a threshold $\tau = 1$ did not produce reliable results. We presume that a single retweet on a topic is not enough of a signal to infer endorsement. Using $\tau = 2$ retweets as threshold proves to be a good trade-off between high selectivity (which hinders analysis) and noise reduction. The resulting size for each retweet graph is listed in Table 2.

In an earlier version of this work [20], when building a conversation graph for a single hashtag, we created an edge between two vertices only if there were "at least two retweets per edge" (in either direction) between the corresponding pair of users. When defining topics as sets of hashtags, there are several ways to generalize this filtering step. The simplest approach considers "two of any" in the set of hashtags that defines the topic. However, this approach is too permissive, and results in an overly-inclusive graph, with spurious relationships and a high level of noise. Instead, we opt to create an edge between two nodes only if there are at least two retweets for any given hashtag between the corresponding pair of users. In other words, the resulting conversation graph for the topic is the union of the retweet graphs for each hashtag in the topic, considered (and filtered) separately.

2. Follow graph. In this approach, we build the follow graph induced by a given hashtag. We select the edges for graph G based on the social connections between Twitter users *who employ the given hashtag*: an edge exists between users u and v if u follows v or vice-versa. We stress that the graph G built with this approach is topic-specific, as the edges in G are constrained to connections between users who discuss the topic that is specified as input to the pipeline.

The rationale for using this graph is based on an assumption of the presence of homophily in the social network, which is a common trait in this setting. To be more precise, we expect that *on a given topic* people will agree more often than not with people they follow, and that for a controversial topic of discussion this phenomenon will be reflected in well-separated partitions of the resulting graph. Note that using the entire social graph would not necessarily produce well-separated partitions that correspond to single topics of discussion, as those partitions would be "blurred" by the existence of additional edges that are due to other reasons (e.g., offline social connections).

On the practical side, while the retweet information is readily available in the stream of tweets, the social network of Twitter is not. Collecting the follower graph thus requires an expensive crawling phase. The resulting graph size for each follow graph is listed in Table 2.

3. Content graph. We create the edges of graph G based on whether users post instances of the same content. Specifically, we experiment with the following three variants: create an edge between two vertices if the users (*i*) use the same hashtag, other than the ones that defines the topic, (*ii*)

[4] We use a hashtag in Russian, #марш which we refer to as #russia_march henceforth, for convenience.

[5] We do not consider 'quote retweets' (retweet with a comment added) in our analysis.

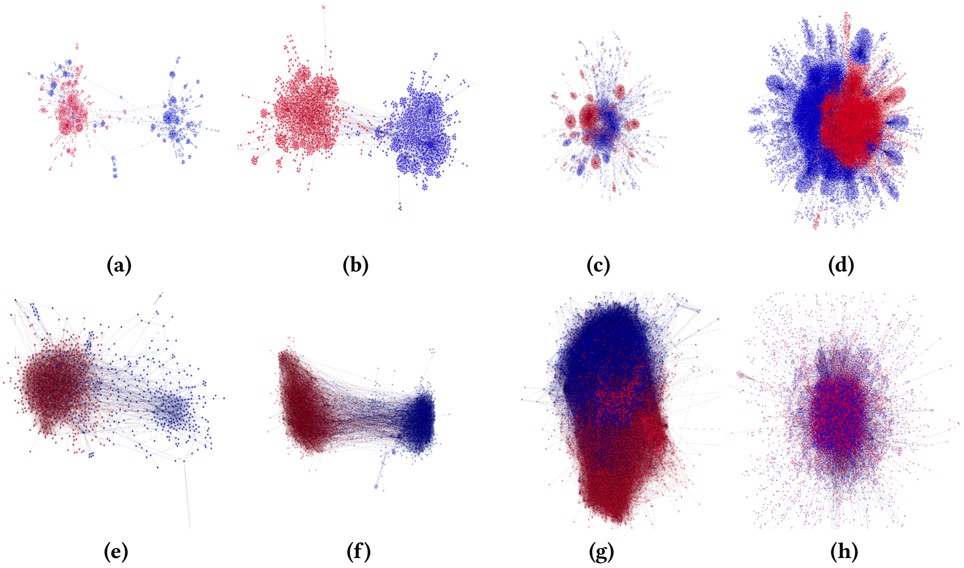


Fig. 3. Sample conversation graphs with retweet (top) and follow (bottom) aspects (visualized using the force-directed layout algorithm in Gephi). The left side is controversial, (a,e) #beefban, (b,f) #russia_march, while the right side is non-controversial, (c,g) #sxsw, (d,h) #germanwings. Only the largest connected component is shown.

share a link to the same URL, or (*iii*) share a link with the same URL domain (e.g., cnn.com is the domain for all pages on the website of CNN).

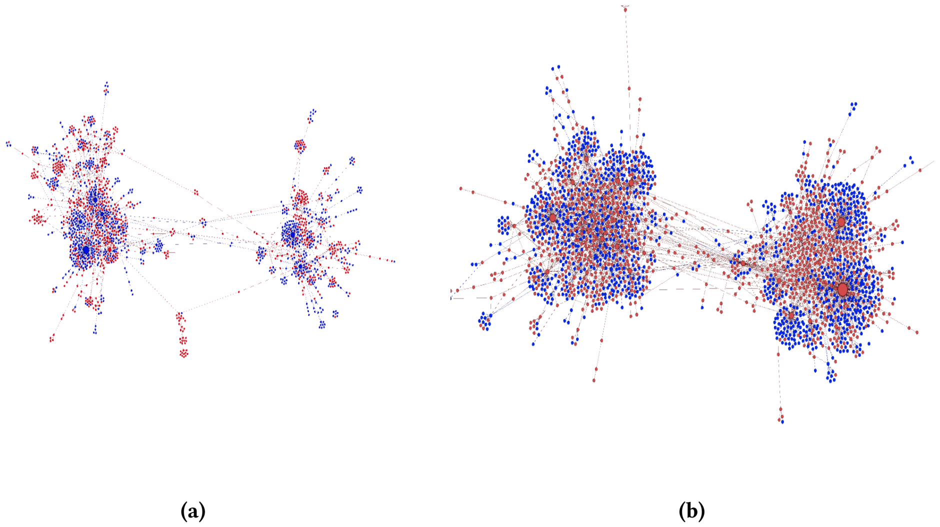
4. Hybrid content & retweet graph. We create edges for graph G according to a state-of-the-art process that blends content and graph information [45]. Concretely, we associate each user with a vector of frequencies of mentions for different hashtags. Subsequently, we create edges between pairs of users whose corresponding vectors have high cosine similarity, and combine them with edges from the retweet graph, built as described above. For details, we refer the interested reader to the original publication [45].

5 GRAPH PARTITIONING

As previously explained, we use a graph partitioning algorithm to produce two partitions on the conversation graph. To do so, we rely on a state-of-the-art off-the-shelf algorithm, METIS [31]. Figure 3 displays the two partitions returned for some of the topics on their corresponding retweet and follow graphs (Figures 3(a)-(d) and Figures 3(e)-(h), respectively).[6] The partitions are depicted in blue or red. The graph layout is produced by Gephi's ForceAtlas2 algorithm [27], and is based solely on the structure of the graph, not on the partitioning computed by METIS. Only the largest connected component is shown in the visualization, though in all the cases the largest connected component makes up > 90% of nodes.

From an initial visual inspection of the partitions identified on retweet and follow graphs, we find that the partitions match well with our intuition of which topics are controversial (the partitions

[6] Other topics show similar trends.


(a)

(b)

Fig. 4. Partitions obtained for (a) #beefban, (b) #russia_march by using the hybrid graph building approach. The partitions are more noisy than those in Figures 3(a,b).

returned by METIS are well separated for controversial topics). To make sure that this initial assessment of the partitions is not an artifact of the visualization algorithm we use, we try other layouts offered by Gephi. In all cases we observe similar patterns. We also manually sample and check tweets from the partitions, to verify the presence of controversy. While this anecdotal evidence is hard to report, indeed the partitions seem to capture the spirit of the controversy.[7]

On the contrary, the partitions identified on content graphs fail to match our intuition. All three variants of the content-based approach lead to sparse graphs and highly overlapping partitions, even in cases of highly controversial issues. The same pattern applies for the hybrid approach, as shown in Figure 4. We also try a variant of the hybrid graph approach with vectors that represent the frequency of different URL domains mentioned by a user, with no better results. We thus do not consider these approaches to graph building any further in the remainder of this paper.

Finally, we try graph partitioning algorithms of other types. Besides METIS (cut based), we test spectral clustering, label propagation, and affiliation-graph-based models. The difference among these methods is not significant, however from visual inspection METIS generates the cleanest partitions.

6 CONTROVERSY MEASURES

This section describes the controversy measures used in this work. For completeness, we describe both those measures proposed by us (§6.1, 6.3, 6.4) as well as the ones from the literature that we use as baselines (§6.5, 6.6).

6.1 Random walk

This measure uses the notion of random walks on graphs. It is based on the rationale that, in a controversial discussion, there are authoritative users on both sides, as evidenced by a large degree

[7] For instance, of these two tweets for #netanyahuspeech from two users on opposing sides, one is clearly supporting the speech https://t.co/OVeWB4XqIg, while the other highlights the negative reactions to it https://t.co/v9RdPudrrC.

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.

Let $G(V,E)$ be the graph built by the first stage and its two partitions X and Y, ($X \cup Y = V$, $X \cap Y = \emptyset$) identified by the second stage of the pipeline. 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. Subsequently, we select one partition at random (each with probability 0.5) and consider a random walk that starts from a random vertex in that partition. The walk terminates when it visits any high-degree vertex (from either side).

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

$$RWC = P_{XX}P_{YY} - P_{YX}P_{XY}, \tag{2}$$

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

$$P_{AB} = Pr[\text{start in partition } A \mid \text{end in partition } B]. \tag{3}$$

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 one when the probability of crossing sides is low, and close to zero when the probability of crossing sides is comparable to that of staying on the same side.

6.2 An efficient variant of the random walk controversy score

The most straightforward way to compute *RWC* is via Monte Carlo sampling. We use this approach in an earlier version of this work [20], with samples of 10 000 random walks. Nevertheless, collecting a large number of samples is computationally intensive, and leads to slow evaluation of *RWC*. In this section, we propose a variant of *RWC* defined as a special case of a *random walk with restart* – thus leading to a much more efficient computation. This variant can handle cases where the random walker gets stuck (i.e., dangling vertices), by using restarts. This feature is important for two reasons: (*i*) retweet graphs (one of our main considerations in this paper) are inherently directed, hence the direction of endorsement should be taken into account, and (*ii*) since these directed graphs are very often star-like, there are a few authoritative users who generate information that spreads through the graph. Our previous Monte Carlo sampling does not take into consideration such graph structure, and the direction of information propagation, as the random walk process needs to be made ergodic for the sampling process to function.

To define the proposed variant of *RWC*, we assume there are two sides for a controversy, defined as two disjoint sets of vertices X and Y. In the original definition of the measure, we start multiple random walks from random vertices on either side, which terminate once they reach a high-degree vertex. For this variant of *RWC*, random walks do not terminate, rather they *restart* once they reach a high-degree vertex.

More formally, we consider two instances of a random walk with restart (RWR), based on whether they start (and restart) from X (start = X) or Y (start = Y). When start = X, the RWR has a restart vector uniformly distributed over X, and zero for vertices in Y (the situation is symmetric for start = Y). Moreover, the random walk runs on a modified graph with all outgoing edges from high-degree vertices removed. This modification transforms the high-degree vertices into dangling vertices, hence forcing the random walk to restart once it reaches one of these vertices.[8]

To formally define this variant of *RWC*, let P_1 and P_2 be the stationary distributions of the RWR obtained for start = X and start = Y, respectively. We consider the conditional probability $Pr[\text{start} = A \mid \text{end} = B^+]$ that the random walk had started on side $A \in \{X, Y\}$, given that at some step at steady-state it is found in one of the high-degree vertices of side $B \in \{X, Y\}$ (denoted as B^+). We thus consider the following four probabilities:

$$P_{X,X^+} = Pr[\text{start} = X \mid \text{end} = X^+] = \frac{\frac{|X|}{|V|}\sum_{v \in X^+} P_1(v)}{\frac{|X|}{|V|}\sum_{v \in X^+} P_1(v) + \frac{|Y|}{|V|}\sum_{v \in X^+} P_2(v)}, \tag{4}$$

$$P_{X,Y^+} = Pr[\text{start} = X \mid \text{end} = Y^+] = \frac{\frac{|X|}{|V|}\sum_{v \in Y^+} P_1(v)}{\frac{|X|}{|V|}\sum_{v \in Y^+} P_1(v) + \frac{|Y|}{|V|}\sum_{v \in Y^+} P_2(v)}, \tag{5}$$

$$P_{Y,Y^+} = Pr[\text{start} = Y \mid \text{end} = Y^+] = \frac{\frac{|Y|}{|V|}\sum_{v \in Y^+} P_2(v)}{\frac{|X|}{|V|}\sum_{v \in Y^+} P_1(v) + \frac{|Y|}{|V|}\sum_{v \in Y^+} P_2(v)}, \tag{6}$$

$$P_{Y,X^+} = Pr[\text{start} = Y \mid \text{end} = X^+] = \frac{\frac{|Y|}{|V|}\sum_{v \in X^+} P_2(v)}{\frac{|X|}{|V|}\sum_{v \in X^+} P_1(v) + \frac{|Y|}{|V|}\sum_{v \in X^+} P_2(v)}. \tag{7}$$

Notice that for the probabilities above we have

$$Pr[\text{start} = X \mid \text{end} = X^+] + Pr[\text{start} = Y \mid \text{end} = X^+] = 1$$

and

$$Pr[\text{start} = X \mid \text{end} = Y^+] + Pr[\text{start} = Y \mid \text{end} = Y^+] = 1$$

as we ought to. The variant of the *RWC* score can be now defined as

$$RWC = P_{XX^+}P_{YY^+} - P_{XY^+}P_{YX^+}, \tag{8}$$

which, like the original version, intuitively captures the difference in the probability of staying on the same side and crossing the boundary.

To verify that the new variant of the score works as expected, we compare it to the original version of the score (obtained via Monte Carlo sampling). The results are shown in Figure 5, from which it can be clearly seen that the new variant is almost identical to the original one. However, for the datasets considered in this work, we found empirically that this algorithm based on random walk with restart is up to 200 times faster compared to the original Monte Carlo algorithm.

6.3 Betweenness

Let us consider the set of edges $C \subseteq E$ in the cut defined by the two partitions X, Y. This measure uses the notion of edge betweenness and how the betweenness of the cut differs from that of the other edges. Note that the cut here refers to the partioning obtained using Metis, as described in Section 3. Recall that the betweenness centrality $bc(e)$ of an edge e is defined as

$$bc(e) = \sum_{s \neq t \in V} \frac{\sigma_{s,t}(e)}{\sigma_{s,t}}, \tag{9}$$

[8] To compute the stationary distribution of the random walks, we use the implementation of Personalized PageRank from NetworkX https://networkx.github.io/documentation/latest/reference/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank.html.

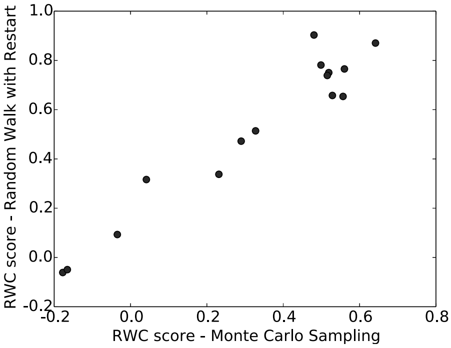


Fig. 5. Comparison between *RWC* scores computed via Monte Carlo sampling and those computed via RWR. Pearson's $r = 0.96$.

where $\sigma_{s,t}$ is the total number of shortest paths between vertices s, t in the graph and $\sigma_{s,t}(e)$ is the number of those shortest paths that include edge e.

The intuition here is that, if the two partitions are well-separated, then the cut will consist of edges that bridge *structural holes* [7]. In this case, the shortest paths that connect vertices of the two partitions will pass through the edges in the cut, leading to high betweenness values for edges in C. On the other hand, if the two partitions are not well separated, then the cut will consist of *strong ties*. In this case, the paths that connect vertices across the two partitions will pass through one of the many edges in the cut, leading to betweenness values for C similar to the rest of the graph.

Given the distributions of edge betweenness on the cut and the rest of the graph, we compute the KL divergence d_{KL} of the two distributions by using kernel density estimation to compute the PDF and sampling 10 000 points from each of these distributions (with replacement). We define the `Betweenness Centrality Controversy` (*BCC*) measure as

$$BCC = 1 - e^{-d_{KL}}, \tag{10}$$

which assumes values close to zero when the divergence is small, and close to one when the divergence is large.

6.4 Embedding

This measure is based on a low-dimensional embedding of graph G produced by Gephi's ForceAtlas2 algorithm [27] (the same algorithm used to produce the plots in Figures 3 and 4). According to Noack [42], a force-directed embedding also maximizes modularity. Based on this observation, the two-dimensional layouts produced by this algorithm indicate a layout with maximum modularity.

Let us consider the two-dimensional embedding $\phi(v)$ of vertices $v \in V$ produced by ForceAtlas2. Given the partition X, Y produced by the second stage of the pipeline, we calculate the following quantities:

- d_X and d_Y, the average embedded distance among pairs of vertices in the same partition, X and Y respectively;
- d_{XY}, the average embedded distance among pairs of vertices across the two partitions X and Y.

Inpsired by the Davies-Bouldin (DB) index [12], we define the `Embedding Controversy` measure *EC* as

$$EC = 1 - \frac{d_X + d_Y}{2d_{XY}}. \tag{11}$$

EC is close to one for controversial topics, corresponding to better-separated graphs and thus to higher degree of controversy, and close to zero for non-controversial topics.

6.5 Boundary Connectivity

This controversy measure was proposed by Guerra et al. [26], and is based on the notion of boundary and internal vertices. Let $u \in X$ be a vertex in partition X; u belongs to the *boundary* of X iff it is connected to at least one vertex of the other partition Y, and it is connected to at least one vertex in partition X that is not connected to any vertex of partition Y. Following this definition, let B_X, B_Y be the set of boundary vertices for each partition, and $B = B_X \cup B_Y$ the set of all boundary vertices. By contrast, vertices $I_X = X - B_X$ are said to be the *internal* vertices of partition X (similarly for I_Y). Let $I = I_X \cup I_Y$ be all internal vertices in either partition. The reasoning for this measure is that, if the two partitions represent two sides of a controversy, then boundary vertices will be more strongly connected to internal vertices than to other boundary vertices of either partition. This intuition is captured in the formula

$$GMCK = \frac{1}{|B|} \sum_{u \in B} \frac{d_i(u)}{d_b(u) + d_i(u)} - 0.5 \tag{12}$$

where $d_i(u)$ is the number of edges between vertex u and internal vertices I, while $d_b(u)$ is the number of edges between vertex u and boundary vertices B. Higher values of the measure correspond to higher degrees of controversy.

6.6 Dipole Moment

This controversy measure was presented by Morales et al. [39], and is based on the notion of *dipole moment* that has its origin in physics. Let $R(u) \in [-1, 1]$ be a polarization value assigned to vertex $u \in V$. Intuitively, extreme values of R (close to -1 or 1) correspond to users who belong most clearly to either side of the controversy. To set the values $R(u)$ we follow the process described in the original paper [39]: we set $R = \pm 1$ for the top-5% highest-degree vertices in each partition X and Y, and set the values for the rest of the vertices by label-propagation. Let n^+ and n^- be the number of vertices V with positive and negative polarization values, respectively, and ΔA the absolute difference of their normalized size $\Delta A = \left|\frac{n^+ - n^-}{|V|}\right|$. Moreover, let gc^+ (gc^-) be the average polarization value among vertices n^+ (n^-) and set d as half their absolute difference, $d = \frac{|gc^+ - gc^-|}{2}$. The dipole moment controversy measure is defined as

$$MBLB = (1 - \Delta A)d. \tag{13}$$

The rationale for this measure is that, if the two partitions X and Y are well separated, then label propagation will assign different extreme (±1) R-values to the two partitions, leading to higher values of the *MBLB* measure. Note also that larger differences in the size of the two partitions (reflected in the value of ΔA) lead to decreased values for the measure, which takes values between zero and one.

7 CONTROVERSY SCORES FOR USERS

The previous sections present measures to quantify the controversy of a conversation graph. In this section, we propose two measures to quantify the controversy 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 controversy score of a user indicates how 'biased' the user is towards a particular side on a topic. For instance, for a topic, say, abortion, pro-choice, pro-life activist groups tweeting consistently about abortion would get a score close to -1/+1 and normal users who interact with both sides 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 controversy score lies exclusively on one side of the graph.

RWC^{user}**:** The first proposed measure is an adaptation of *RWC*. As input, we are given a user $u \in V$ in the graph and a partitioning of the graph into two sides, defined as disjoint sets of vertices X and Y. We then consider a random walk that starts – and restarts – at the given user u. Moreover, as with *RWC*, the high-degree vertices on each side (X^+ and Y^+) are treated as dangling vertices – whenever the random walk reaches these vertices, it teleports to vertex u with probability 1 in the next step. To quantify the controversy of u, we ask how often the random walk is found on vertices that belong to either side of the controversy. Specifically, for each user u, we consider the conditional probabilities $Pr[\text{start} = u \mid \text{end} = X^+]$ and $Pr[\text{start} = u \mid \text{end} = Y^+]$, we estimate them by using the power iteration method. Assuming that user u belongs to side X of the controversy (i.e., $u \in X$), their controversy is defined as:

$$RWC^{user}(u,X) = \frac{Pr[\text{start} = u \mid \text{end} = X^+]}{Pr[\text{start} = u \mid \text{end} = X^+] + Pr[\text{start} = u \mid \text{end} = Y^+]}. \tag{14}$$

Expected hitting time: The second proposed measure is also random-walk-based, but defined on the expected number of steps to hit the high-degree vertices on either side. 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^+ or Y^+). 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 controversy score of a user is then defined as

$$\rho(u) = \rho^X(u) - \rho^Y(u) \in (-1, 1). \tag{15}$$

Following the definition, a vertex that, compared to most other vertices, is very close to high-degree vertices X^+ will have $\rho^X(u) \approx 1$; and if the same vertex is very far from high-degree vertices Y^+, we'll have $\rho^Y(u) \approx 0$ – leading to a controversy score $\rho(u) \approx 1 - 0 = 1$. The opposite is true for vertices that are far from X^+ but close to Y^+ – leading to a controversy score $\rho(u) \approx -1$.

7.1 Comparison with BiasWatch

BiasWatch [36] is a recently-proposed, light-weight approach to compute controversy scores for users on Twitter. At a high level, the BiasWatch approach consists of the following steps:

1. Hand pick a small set of seed hashtags to characterize the two sides of a controversy (e.g., #prochoice vs. #prolife);
2. Expand the seed set of hashtags based on co-occurrence;
3. Use the two sets of hashtags, identify strong partisans in the graph (users with high controversy score);
4. Assign controversy scores to other users via a simple label propagation approach.

We compare the controversy scores obtained by our approaches to the ones obtained by BiasWatach[9] on two sets of datasets: tweets matching the hashtags (*i*) #obamacare, #guncontrol, and #abortion, provided by Lu et al. [36] and (*ii*) the datasets in Table 2. We compute the Pearson correlation between our measure based on Expected hitting time and BiasWatch; the results are shown in Figure 6. We omit the comparison with RWC^{user} scores as they are almost identical to the ones by BiasWatch.

The authors also provide datasets which contain human annotations for controversy score (in the range [-2,2]) for 500 randomly selected users. We discretize our controversy scores to the same range, and compute the 5-category Fleiss' κ value. The κ value is 0.35, which represents a 'fair' level of agreement, according to Landis and Koch [35].

Topic	Pearson correlation
#abortion	0.51
#obamacare	0.48
#guncontrol	0.42
#beefban	0.41
#baltimoreriots	0.41
#netanyahuspeech	0.41
#nemtsov	0.38
#indiana	0.40
#indiasdaughter	0.39
#ukraine	0.40

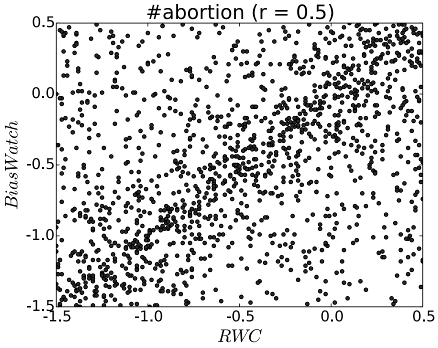


Fig. 6. (left) Pearson's r between the scores obtained by our algorithm and BiasWatch. (right) Sample scatter plot for #abortion.

Our approach thus provides results that are similar to the state-of-the-art approach. Our method also has two advantages over the BiasWatch measure: (i) Even though we do not make use of any content information in our measure, we perform at par; and (ii) RWC^{user} provides an intuitive extension to our RWC measure. Given this unified framework, it is possible to design ways to reduce controversy, e.g. by connecting opposing views [21, 22], and such a unified formulation can help us define principled objective functions to approach these tasks.

8 EXPERIMENTS

In this section we report the results of the various configurations of the pipeline proposed in this paper. As previously stated, we omit results for the content and hybrid graph building approaches presented in Section 4, as they do not perform well. We instead focus on the retweet and follow graphs, and test all the measures presented in Section 6 on the topics described in Table 2. In addition, we test all the measures on a set of external datasets used in previous studies [1, 10, 26] to validate the measures against a known ground truth. Finally, we use an evolving dataset from Twitter collected around the death of Venezuelan president Hugo Chavez [39] to show the *evolution* of the controversy measures in response to high-impact events.

To avoid potential overfitting, we use only eight graphs as testbed during the development of the measures, half of them controversial (beefban, nemtsov, netanyahu, russia_march) and half

[9] For BiasWatch we use parameters $\mu_1 = 0.1$, $\mu_2 = 0.4$, optimization method 'COBYLA', cosine similarity threshold 0.4, and 10 nearest neighbors for hashtag extension.

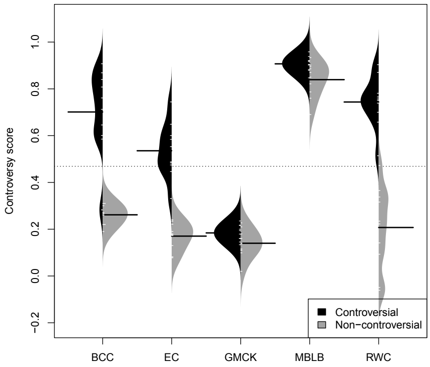


Fig. 7. Controversy scores on *retweet* graphs of various controversial and non-controversial datasets.

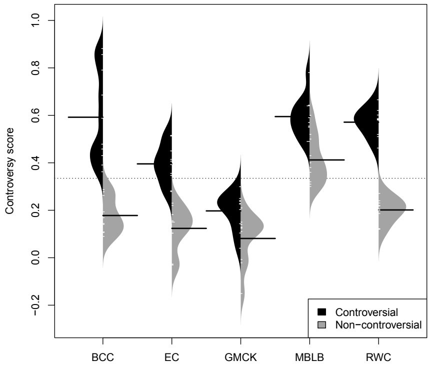


Fig. 8. Controversy scores on *follow* graphs of various controversial and non-controversial datasets.

Table 3. Results on external datasets. The 'C?' column indicates whether the previous study considered the dataset controversial (ground truth).

Dataset	$\|V\|$	$\|E\|$	C?	*RWC*	*BCC*	*EC*	*GMCK*	*MBLB*
Political blogs	1222	16 714	✓	0.42	0.53	0.49	0.18	0.45
Twitter politics	18 470	48 053	✓	0.77	0.79	0.62	0.28	0.34
Gun control	33 254	349 782	✓	0.70	0.68	0.55	0.24	0.81
Brazil soccer	20 594	82 421	✓	0.67	0.48	0.68	0.17	0.75
Karate club	34	78	✓	0.11	0.64	0.51	0.17	0.11
Facebook university	281	4389	✗	0.35	0.26	0.38	0.01	0.27
NYC teams	95 924	176 249	✗	0.34	0.24	0.17	0.01	0.19

non-controversial (sxsw, germanwings, onedirection, ultralive). This procedure resembles a 40/60% train/test split in traditional machine learning applications.[10]

8.1 Twitter hashtags

Figure 7 and Figure 8 report the scores computed by each measure for each of the 20 hashtags, on the retweet and follow graph, respectively. Each figure shows a set of beanplots,[11] one for each measure. Each beanplot shows the estimated probability density function for a measure computed on the topics, the individual observations are shown as small white lines in a one-dimensional scatter plot, and the median as a longer black line. The beanplot is divided into two groups, one for controversial topics (left/dark) and one for non-controversial ones (right/light). A larger separation of the two distributions indicates that the measure is better at capturing the characteristics of controversial topics. For instance, this separation is fundamental when using the controversy score as a feature in a classification algorithm.

Figures 7 and 8 clearly show that *RWC* is the best measure on our datasets. *BCC* and *EC* show varying degrees of separation and overlap, although *EC* performs slightly better as the distributions

[10] A demo of our controversy measures can be found at https://users.ics.aalto.fi/kiran/controversy.

[11] A beanplot is an alternative to the boxplot for visual comparison of univariate data among groups.

are more concentrated, while *BCC* has a very wide distribution. The two baselines *GMCK* and *MBLB* instead fail to separate the two groups. Especially on the retweet graph, the two groups are almost indistinguishable.

For all measures the median score of controversial topics is higher than for non-controversial ones. This result suggests that both graph building methods, retweet and follow, are able to capture the difference between controversial and non-controversial topics. Given the broad range of provenience of the topics covered by the dataset, and their different characteristics, the consistency of the results is very encouraging.

8.2 External datasets

We have shown that our approach works well on a number of datasets extracted in-the-wild from Twitter. But, how well does it generalize to datasets from different domains?

We obtain a comprehensive group of datasets kindly shared by authors of previous works: *Political blogs*, links between blogs discussing politics in the US [1]; *Twitter politics*, Twitter messages pertaining to the 2010 midterm election in US [10]; and the following five graphs used in the study that introduced *GMCK* [26], (a) *Gun control*, retweets about gun control after the shooting at the Sandy Hook school; (b) *Brazil soccer*, retweets about to two popular soccer teams in Brazil; (c) *Karate club*, the well-known social network by [49]; (d) *Facebook university*, a social graph among students and professors at a Brazilian university; (e) *NYC teams*, retweets about two New York City sports teams.

Table 3 shows a comparison of the controversy measures under study on the aforementioned datasets.[12] For each dataset we also report whether it was considered controversial in the original paper, which provides a sort of “ground truth” to evaluate the measures against.

All the measures are able to distinguish controversial graphs to some extent, in the sense that they return higher values for the controversial cases. The only exception is Karate club. Both *RWC* and *MBLB* report low controversy scores for this graph. It is possible that the graph is too small for such random-walk-based measures to function properly. Conversely, *BCC* is able to capture the desired behavior, which suggests that shortest-path and random-walk based measures might have a complementary function.

Interestingly, while the Political blogs datasets is often considered a gold standard for polarization and division in online political discussions, all the measures agree that it presents only a moderate level of controversy. Conversely, the Twitter politics dataset is clearly one of the most controversial one across all measures. This difference suggests that the measures are more geared towards capturing the dynamics of controversy as it unfolds on social media, which might differ from more traditional blogs. For instance, one such difference is the *cost* of an endorsement: placing a link on a blog post arguably consumes more mental resources than clicking on the retweet button.

For the ‘Gun control’ dataset, Guerra et al. need to manually distinguish three different partitions in the graph: gun rights advocates, gun control supporters, and moderates. Our pipeline is able to find the two communities with opposing views (grouping together gun control supporters and moderates, as suggested in the original study) without any external help. All measures agree with the conclusions drawn in the original paper that this topic is highly controversial.

Note that even though from the results in Table 3, RWC, BCC and EC appear to outperform each other, it is not the case. These methods are not comparable, meaning, a score of 0.5 for RWC is not the same as a 0.5 for BCC. The insight we can draw from these results is that our methods are able

[12] The datasets provided by Guerra et al. [26] are slightly different from the ones used in the original paper because of some irreproducible filtering used by the authors. We use the datasets provided to us verbatim.

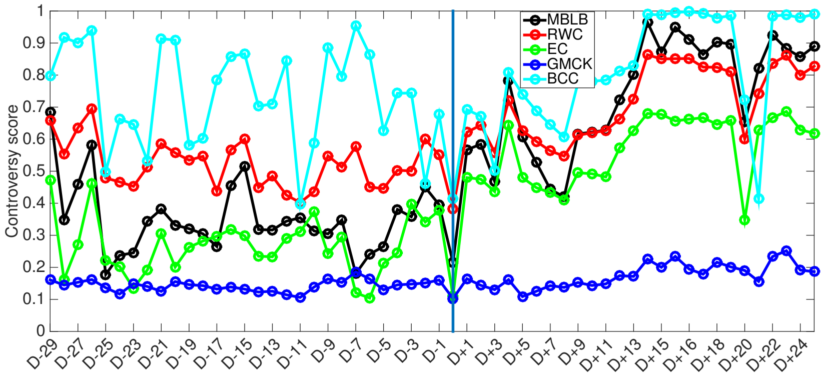


Fig. 9. Controversy scores on 56 retweet graphs from Morales et al. Day 'D' (indicated by the blue vertical line) indicates the announcement of the death of president Hugo Chavez.

to identify a controversial topic from a non-controversial topic consistently, irrespective of the domain and are able to do it better than existing methods (GMCK and MBLB).

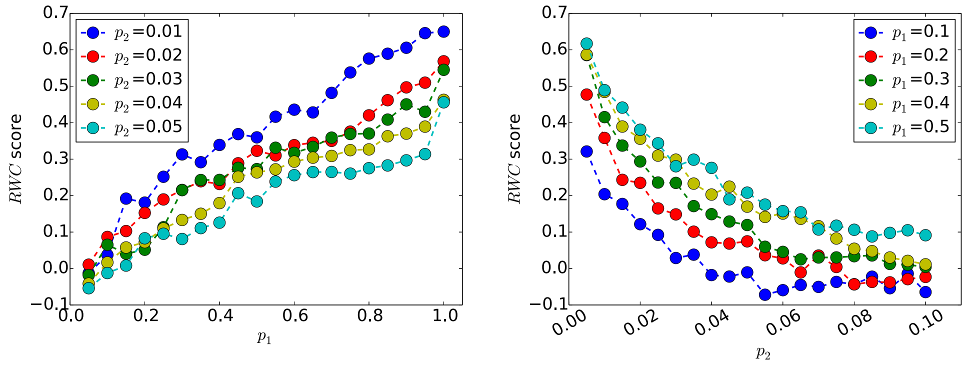
8.3 Evolving controversy

We have shown that our approach also generalizes well to datasets from different domains. But in a real deployment the measures need to be computed continuously, as new data arrives. How well does our method work in such a setting? And how do the controversy measures evolve in response to high-impact events?

To answer these questions, we use a dataset from the study that introduced *MBLB* [39]. The dataset comprises Twitter messages pertaining to political events in Venezuela around the time of the death of Hugo Chavez (Feb-May 2013). The authors built a retweet graph for each of the 56 days around the day of the death (one graph per day).

Figure 9 shows how the intensity of controversy evolves according to the measures under study (which occurs on day 'D'). The measure proposed in the original paper, *MBLB*, which we use as 'ground truth', shows a clear decrease of controversy on the day of the death, followed by a progressive increase in the controversy of the conversation. The original interpretation states that on the day of the death a large amount of people, also from other countries, retweeted news of the event, creating a single global community that got together at the shock of the news. After the death, the ruling and opposition party entered in a fiery discussion over the next elections, which increased the controversy.

All the measures proposed in this work show the same trend as *MBLB*. Both *RWC* and *EC* follow very closely the original measure (Pearson correlation coefficients r of 0.944 and 0.949, respectively), while *BCC* shows a more jagged behavior in the first half of the plot ($r = 0.743$), due to the discrete nature of shortest paths. All measures however present a dip on day 'D', an increase in controversy in the second half, and another dip on day 'D+20'. Conversely, *GMCK* reports an almost constant moderate value of controversy during the whole period ($r = 0.542$), with barely noticeable peaks and dips. We conclude that our measures generalize well also to the case of evolving graphs, and behave as expected in response to high-impact events.



(a)

(b)

Fig. 10. *RWC* scores for synthetic Erdös-Rényi graphs planted with two communities. p_1 is the intra-community edge probability, while p_2 is the inter-community edge probability.

8.4 Simulations

Given that *RWC* is the best-performing score among the ones in this study, we focus our attention solely on it henceforth. To measure the robustness of the *RWC* score, we generate random Erdös-Rényi graphs with varying community structure, and compute the *RWC* score on them. Specifically, to mimic community structure, we plant two separate communities with intra-community edge probability p_1. That is, p_1 defines how dense these communities are within themselves. We then add random edges between these two communities with probability p_2. Therefore, p_2 defines how connected the two communities are. A higher value of p_1 and a lower value of p_2 create a clearer two-community structure.

Figure 10 shows the *RWC* score for random graphs of 2000 vertices for two different settings: plotting the score as a function of p_1 while fixing p_2 (Figure 10a), and vice-versa (Figure 10b). The *RWC* score reported is the average over ten runs. We observe a clear pattern: the *RWC* score increases as we increase the density within the communities, and decreases as we add noise to the community structure. The effects of the parameters is also expected, for a given value of p_1, a smaller value of p_2 generates a larger *RWC* score, as the communities are more well separated. Conversely, for a given value of p_2, a larger value of p_1 generates a larger *RWC* scores, as the communities are denser.

8.5 Controversy detection in the wild

In most of the experiments presented so far, we hand-picked known topics which are controversial and show that our method is able to separate them from the non-controversial topics. To check whether our system works in a real-world setting, we deploy it in the wild to explore actual topics of discussion on Twitter and detect the ones that are controversial. More specifically, we obtain daily trending hashtags (both US and worldwide) on the platform for a period of three months (June 25 – September 19, 2015). Then, we obtain all tweets that use these hashtags, and create retweet graphs (as described in Section 4). Finally, we apply the *RWC* measure on these conversation graphs to identify controversial hashtags.

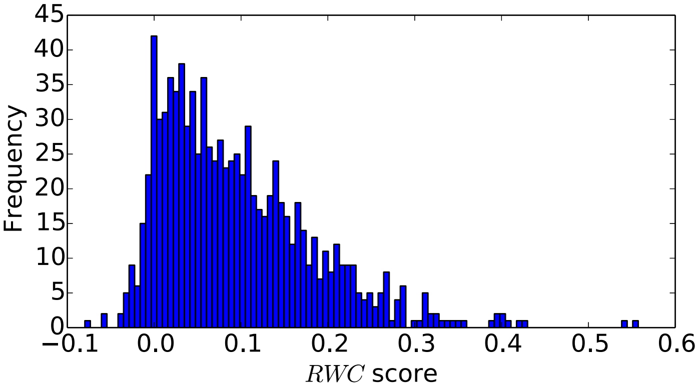


Fig. 11. Frequency of *RWC* scores for hashtags trending from June to September 2015.

The results can be explored in our online demo [19].[13] To mention a few examples, our system was able to identify the following controversial hashtags:

- #whosiburningblackchurches (score 0.332): A hashtag about the burning of predominantly black churches.[14]
- #communityshield (score 0.314): Discussion between the fans of two sides of a soccer game.[15]
- #nationalfriedchickenday (score 0.393): A debate between meat lovers and vegetarians about the ethics of eating meat.

Moreover, based on our experience with our system, most hashtags that are reported as trending on Twitter concern topics that are not controversial. Figure 11 shows the histogram of the *RWC* score over the 924 trending hashtags we collected. A majority of these hashtags have an *RWC* score around zero.

9 CONTENT

In this section we explore alternative approaches to measuring controversy that use only the content of the discussion rather than the structure of user interactions. As such, these methods do not fit in the pipeline described in Section 3. The question we address is "does content help in measuring the controversy of a topic?" In particular, we test two types of features extracted from the content. The first, is a typical IR-inspired bag-of-words representation. The second instead is based on NLP tools for sentiment analysis.

9.1 Bag of words

We take as input the raw content of the social media posts, in our case the tweets pertaining to a specific topic. We represent each tweet as a vector in a high-dimensional space composed of the words used in the whole topic, after standard preprocessing used in IR (lowercasing, stopword

[13] https://users.ics.aalto.fi/kiran/controversy/table.php

[14] https://erlc.com/article/explainer-whoisburningblackchurches.

[15] https://en.wikipedia.org/wiki/2015_FA_Community_Shield.

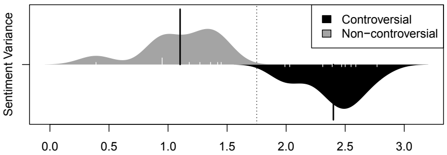


Fig. 12. Sentiment variance controversy score for controversial and non-controversial topics.

removal, stemming). Following the lines of our main pipeline, we group these vectors in two clusters by using CLUTO [30] with cosine distance.

The underlying assumption is that the two sides, while sharing the use of the hashtag for the topic, use different vocabularies in reference to the issue at hand. For example, for #beefban a side may be calling for "freedom" while the opposing one for "respect." We use KL divergence as a measure of distance between the vocabularies of the two clusters, and the I2 measure [37] of clustering heterogeneity.

We use an unpaired Wilcoxon rank-sum test at the $p = 0.05$ significance level, but we are unable to reject the null hypothesis that there is no difference in these measures between the controversial and non-controversial topics. Therefore, there is not enough signal in the content representation to discern between controversial and non-controversial topics with confidence. This result suggests that the bag-of-words representation of content is not a good basis for our task. It also agrees with our earlier attempts to use content to build the graph used in the pipeline (see Section 4) – which suggests that using content for the task of quantifying controversy might not be straightforward.

9.2 Sentiment analysis

Next, we resort to NLP techniques for sentiment analysis to analyze the content of the discussion. We use SentiStrength [47] trained on tweets to give a sentiment score in $[-4, 4]$ to each tweet for a given topic. In this case we do not try to cluster tweets by their sentiment. Rather, we analyze the difference in distribution of sentiment between controversial and non-controversial topics.

While it is not possible to say that controversial topics are more positive or negative than non-controversial ones, we can detect a difference in their variance. Indeed, controversial topics have a higher variance than non-controversial ones, as shown in Figure 12. Controversial ones have a variance of at least 2, while non-controversial ones have a variance of at most 1.5.

In practice, the "tones" with which controversial topics are debated are stronger, and sentiment analysis is able to detect this aspect. While this signal is clear, it is not straightforward to incorporate it into the measures based on graph structure. Moreover, this feature relies on technologies that do not work reliably for languages other than English and hence cannot be applied for topics such as #russia_march.

10 DISCUSSION

The task we tackle in this work is certainly not an easy one, and this study has some limitations, which we discuss in this section. We also report a set of negative results that we produced while coming up with the measures presented. We believe these results will be very useful in steering this research topic towards a fruitful direction. Table 4 provides a summary of the various graph building strategies and controversy measures we tried for quantifying controversy.

Table 4. Summary of various graph building and controversy measures tried. * indicates the methods that worked.

Graphs	Retweet* Follow* Content Mention Hybrid (content + retweet, mention + retweet)
Measures	Random Walk* Edge betweenness* Embedding Boundary Connectivity Dipole Moment Cut-based measures (conductance, cut ratio) Sentiment analysis* Modularity SPID

10.1 Limitations

Twitter only. We present our findings mostly on datasets coming from Twitter. 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 measures generalize well to datasets from other social media and the Web.

Choice of data. We manually pick the controversial topics in our dataset, which might introduce bias. In our choice we represent a broad set of typical controversial issues coming from religious, societal, racial, and political domains. Unfortunately, ground truths for controversial topics are hard to find, especially for ephemeral issues. However, the topics are unanimously judged controversial by the authors. Moreover, the hashtags represent the intuitive notion of controversy that we strive to capture, so human judgement is an important ingredient we want to use.

Overfitting. While this work presents the largest systematic study on controversy in social media so far, we use only 20 topics for our main experiment. Given the small number of examples, the risk of overfitting our measures to the dataset is real. We reduce this risk by using only 40% of the topics during the development of the measures. Additionally, our measures agree with previous independent results on external datasets, which further decreases the likelihood of overfitting.

Reliance on graph partitioning. Our pipeline relies on a graph partitioning stage, whose quality is fundamental for the proper functioning of the controversy measures. Given that graph partitioning is a hard but well studied problem, we rely on off-the-shelf techniques for this step. A measure that bypasses this step entirely is highly desirable, and we report a few unsuccessful attempts in the next subsection.

Multisided controversies. Not all controversies involve only two sides with opposing views. Some times discussions are multifaceted, or there are three or more competing views on the field. The principles behind our measures neatly generalize to multisided controversies. However, in this case the graph partitioning component needs to automatically find the optimal number of partitions. We defer experimental study of such cases to an extended version of this paper.

Evaluation. Defining what is controversial/polarized can be subjective. There are many ways to define what is controversial, depending on the context, subject and field of study, e.g. See [6] for around a dozen ways to define polarization. Our evaluation is based on our intuitive labelling that a topic is controversial/polarized. This might not always be true, but given that the alternative is to hand-label/survey the thousands of users, we presume that this assumption is reasonable for developing methods that can be adapted to large scale systems.

10.2 Negative results

We briefly review a list of methods that failed to produce reliable results and were discarded early in the process of refining our controversy measures.

Mentions graph. Conover et al. [10] rely on the mention graph in Twitter to detect controversies. However, in our dataset the mention graphs are extremely sparse given that we focus on short-lived events. Merging the mentions into the retweet graph does not provide any noticeable improvement.

Previous studies have also shown that people retweet similar ideologies but mention across ideologies [5]. We exploit this intuition by using correlation clustering for graph partitioning, with negative edges for mentions. Alas, the results are qualitatively worse than those obtained by METIS.

Cuts. Simple measures such as size of the cut of the partitions do not generalize across different graphs. Conductance (in all its variants) also yields poor results. Prior work identifies controversies by comparing the structure of the graph with randomly permuted ones [10]. Unfortunately, we obtain equally poor results by using the difference in conductance with cuts obtained by METIS and by random partitions.

Community structure. Good community structure in the conversation graph is often understood as a sign that the graph is polarized or controversial. However, this is not always the case. We find that both assortativity and modularity (which have been previously used to identify controversy) do not correlate with the controversy scores, and are not good predictors for how controversial a topic is. The work by Guerra, et al [26] presents clear arguments and examples of why modularity should be avoided.

Partitioning. As already mentioned, bypassing the graph partitioning to compute the measure is desirable. We explore the use of the all pairs expected hitting time computed by using SimRank [29]. We compute the SPID (ratio of variance to mean) of this distribution, however results are mixed.

10.3 Conclusions

In this paper, we performed the first large-scale systematic study for quantifying controversy in social media. We have shown that previously-used measures are not reliable and demonstrated that controversy can be identified both in the retweet and topic-induced follow graph. We have also shown that simple content-based representations do not work in general, while sentiment analysis offers promising results.

Among the measures we studied, the random-walk-based *RWC* most neatly separates controversial topics from non-controversial ones. Besides, our measures gracefully generalize to datasets from other domains and previous studies.

This work opens several avenues for future research. First, it is worth exploring alternative approaches and testing additional features, such as, following a generative-model-based approach, or exploiting the temporal evolution of the discussion of a topic.

From the application point of view, the controversy score can be used to generate recommendations that foster a healthier "news diet" on social media. Given the ever increasing impact of polarizing figures in our daily politics and the rise in polarization in the society [13, 23], it is important to not restrict ourselves to our own 'bubbles' or 'echo chambers' [43, 46]. Our methods for identifying controversial topics can be used as building blocks for designing such systems to reduce controversy on social media [21, 22] by connecting social media users with content outside their own bubbles.

In addition, polarization by itself may not be a bad thing. Many studies [11, 41] argue that a democracy needs deliberation and polarization/controversy 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 to what extent a discussion is polarized, so that things do not spiral out of control, and create isolated echo chambers. Our paper tries to contribute methods that help in this setting, by measuring the degree of polarization of a topic.

Acknowledgements. This work has been supported by the Academy of Finland project "Nestor" (286211) and the EC H2020 RIA project "SoBigData" (654024).

REFERENCES

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

[2] Leman Akoglu. 2014. Quantifying Political Polarity Based on Bipartite Opinion Networks. In *ICWSM*.

[3] Md Tanvir Al Amin, Charu Aggarwal, Shuochao Yao, Tarek Abdelzaher, and Lance Kaplan. 2017. *Unveiling Polarization in Social Networks: A Matrix Factorization Approach.* Technical Report. IEEE.

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

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

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

[7] Ronald S Burt. 2009. *Structural holes: The social structure of competition.* Harvard university press.

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

[9] Mauro Coletto, Kiran Garimella, Aristides Gionis, and Claudio Lucchese. 2017. A Motif-based Approach for Identifying Controversy. In *Proceedings of the 10th International on Conference on Web and Social Media*. AAAI.

[10] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. In *ICWSM*.

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

[12] David L. Davies and Donald W. Bouldin. 1979. A Cluster Separation Measure. *IEEE TPAMI* 1, 2 (1979), 224–227.

[13] Michael Dimock, Carroll Doherty, Jocelyn Kiley, and Russ Oates. 2014. Political polarization in the American public: How increasing ideological uniformity and partisan antipathy affect politics, compromise and everyday life. *Washington, DC: Pew Research Center* (2014).

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

[15] Abraham Doris-Down, Husayn Versee, and Eric Gilbert. 2013. Political blend: an application designed to bring people together based on political differences. In *C&T*. 120–130.

[16] Kevin M Esterling, Archon Fung, and Taeku Lee. 2010. How Much Disagreement is Good for Democratic Deliberation? The CaliforniaSpeaks Health Care Reform Experiment. *SSRN* (2010).

[17] Wei Feng, Jiawei Han, Jianyong Wang, Charu Aggarwal, and Jianbin Huang. 2015. STREAMCUBE: Hierarchical Spatio-temporal Hashtag Clustering for Event Exploration over the Twitter Stream. In *ICDE*.

[18] Seth R Flaxman, Sharad Goel, and Justin M Rao. 2015. Filter Bubbles, Echo Chambers, and Online News Consumption. (2015).

[19] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016a. Exploring Controversy in Twitter. In *CSCW [demo]*. 33–36.

[20] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016b. Quantifying Controversy in Social Media. In *WSDM*. 33–42.

[21] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017a. Factors in Recommending Contrarian Content on Social Media. In *WebSci*.

[22] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017b. Reducing Controversy by Connecting Opposing Views. In *WSDM*. 81–90.

[23] Kiran Garimella and Ingmar Weber. 2017. A Long-Term Analysis of Polarization on Twitter. In *ICWSM*.

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

[25] Catherine Grevet, Loren G Terveen, and Eric Gilbert. 2014. Managing political differences in social media. In *CSCW*. 1400–1408.

[26] Pedro Henrique Calais Guerra, Wagner Meira Jr, Claire Cardie, and Robert Kleinberg. 2013. A Measure of Polarization on Social Media Networks Based on Community Boundaries. In *ICWSM*.

[27] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. (2014).

[28] Myungha Jang, John Foley, Shiri Dori-Hacohen, and James Allan. 2016. Probabilistic Approaches to Controversy Detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2069–2072.

[29] Glen Jeh and Jennifer Widom. 2002. SimRank: A measure of structural-context similarity. In *KDD*. 538–543.

[30] George Karypis. 2002. CLUTO - A clustering toolkit. (2002).

[31] George Karypis and Vipin Kumar. 1995. METIS - Unstructured Graph Partitioning and Sparse Matrix Ordering System. (1995).

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

[33] Juhi Kulshrestha, Muhammad Bilal Zafar, Lisette Espin Noboa, Krishna P Gummadi, and Saptarshi Ghosh. 2015. Characterizing Information Diets of Social Media Users. In *ICWSM*.

[34] Michael LaCour. 2012. A balanced news diet, not selective exposure: Evidence from a direct measure of media exposure. *SSRN* (2012).

[35] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[36] Haokai Lu, James Caverlee, and Wei Niu. 2015. BiasWatch: A Lightweight System for Discovering and Tracking Topic-Sensitive Opinion Bias in Social Media. In *CIKM*. 213–222.

[37] Ujjwal Maulik and Sanghamitra Bandyopadhyay. 2002. Performance evaluation of some clustering algorithms and validity indices. *IEEE TPAMI* 24, 12 (2002), 1650–1654.

[38] Yelena Mejova, Amy X Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and Sentiment in Online News. *arXiv preprint arXiv:1409.8152* (2014).

[39] AJ Morales, J Borondo, JC Losada, and RM Benito. 2015. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos* 25, 3 (2015).

[40] Sean A Munson, Stephanie Y Lee, and Paul Resnick. 2013. Encouraging Reading of Diverse Political Viewpoints with a Browser Widget.. In *ICWSM*.

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

[42] Andreas Noack. 2009. Modularity clustering is force-directed layout. *Physical Review E* 79, 2 (2009), 026102.

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

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

[45] Yiye Ruan, David Fuhry, and Srinivasan Parthasarathy. 2013. Efficient community detection in large networks using content and links. In *WWW*. 1089–1098.

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

[47] Mike Thelwall. 2013. Heart and soul: Sentiment strength detection in the social web with SentiStrength. In *CyberEmotions*. 1–14.

[48] Mikalai Tsytsarau, Themis Palpanas, and Kerstin Denecke. 2011. Scalable detection of sentiment-based contradictions. *DiversiWeb, WWW* 2011 (2011).

[49] Wayne Zachary. 1977. An Information Flow Model for Conflict and Fission in Small Groups. *J. of Anthropological Research* 33 (1977), 452–473.

Publication 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.

Balancing information exposure in social networks

Kiran Garimella
Aalto University & HIIT
Helsinki, Finland
kiran.garimella@aalto.fi

Aristides Gionis
Aalto University & HIIT
Helsinki, Finland
aristides.gionis@aalto.fi

Nikos Parotsidis
University of Rome Tor Vergata
Rome, Italy
nikos.parotsidis@uniroma2.it

Nikolaj Tatti
Aalto University & HIIT
Helsinki, Finland
nikolaj.tatti@aalto.fi

Abstract

Social media has brought a revolution on how people are consuming news. Beyond the undoubtedly large number of advantages brought by social-media platforms, a point of criticism has been the creation of *echo chambers* and *filter bubbles*, caused by *social homophily* and *algorithmic personalization*.

In this paper we address the problem of *balancing the information exposure* in a social network. We assume that two opposing campaigns (or viewpoints) are present in the network, and that network nodes have different preferences towards these campaigns. Our goal is to find two sets of nodes to employ in the respective campaigns, so that the overall information exposure for the two campaigns is *balanced*. We formally define the problem, characterize its hardness, develop approximation algorithms, and present experimental evaluation results.

Our model is inspired by the literature on *influence maximization*, but there are significant differences from the standard model. First, *balance* of information exposure is modeled by a *symmetric difference* function, which is neither monotone nor submodular, and thus, not amenable to existing approaches. Second, while previous papers consider a setting with selfish agents and provide bounds on *best-response* strategies (i.e., move of the last player), we consider a setting with a centralized agent and provide bounds for a global objective function.

1 Introduction

Social-media platforms have revolutionized many aspects of human culture, among others, the way people are exposed to information. A recent survey estimates that 62% of adults in the US get their news on social media [15]. Despite providing many desirable features, such as, searching, personalization, and recommendations, one point of criticism is that social media amplify *echo chambers* and *filter bubbles*: users get less exposure to conflicting viewpoints and are isolated in their own informational bubble. This phenomenon is contributed to social homophily and algorithmic personalization, and is more acute for controversial topics [9, 12, 14].

In this paper we address the problem of reducing the filter-bubble effect by balancing information exposure among users. We consider social-media discussions around a topic that are characterized by two or more *conflicting viewpoints*. Let us refer to these viewpoints as *campaigns*. Our approach follows the popular paradigm of influence propagation [18]: 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

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

problem of finding an optimal *selfish strategy* for each campaign separately. Instead we consider a *centralized agent* responsible for balancing information exposure for the two campaigns Consider the following motivating examples.

Example 1: Social-media companies 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 legislation.[1] Consider a controversial topic being discussed in social-media platform X, which has led to polarization and filter bubbles. As part of a new filter-bubble bursting service, platform X would like to disseminate two high-quality and thought-provoking dueling op-eds, articles, one for each side, which 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 the technical standpoint, we consider the following problem setting: We assume that information is propagated in the network according to the *independent-cascade model* [18]. 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 occurs according to some information-propagation model.

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).

We show that the problem of balancing the information exposure is $\mathbf{NP}$-hard. We develop different approximation algorithms for the different settings we consider, as well as heuristic variants of the proposed algorithm. We experimentally evaluate our methods, on several real-world datasets.

Although our approach is inspired by the large body of work on information propagation, and resembles previous problem formulations for competitive viral marketing, there are significant differences. In particular:

- This is the first paper to address the problem of *balancing information exposure* and *breaking filter bubbles*, 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.

Omitted proofs, figures, and tables are provided as supplementary material. Moreover, our datasets and implementations are publicly available.[2]

2 Related Work

Detecting and breaking filter bubbles. Several studies have observed that users in online social networks prefer to associate with like-minded individuals and consume agreeable content. This phenomenon leads to *filter bubbles*, *echo chambers* [25], and to online polarization [1, 9, 12, 22].

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

[2] `https://users.ics.aalto.fi/kiran/BalanceExposure/`

Once these filter bubbles are detected, the next step is to try to overcome them. One way to achieve this is by making recommendations to individuals of opposing viewpoints. This idea has been explored, in different ways, by a number of studies in the literature [13, 19]. However, previous studies address the problem of breaking filter bubbles by the means of *content recommendation*. To the best of our knowledge, this is the first paper that considers an *information diffusion* approach.

Information diffusion. Following a large body of work, we model diffusion using the *independent-cascade model* [18]. 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 [23] 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. Carnes et al. [7] propose two models for competitive diffusion. Subsequently, several other models have been proposed [4, 10, 11, 17, 21, 27, 28].

Most of the work on *competitive information diffusion* consider the problem of selecting the best k seeds for one campaign, for a given objective, in the presence of competing campaigns [3, 6]. Bharathi et al. [3] 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 [2, 16, 26].

The work that is most related to ours, in the sense of considering a *centralized authority*, is the one by Borodin et al. [5]. 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. Nevertheless, we can use this framework as a heuristic baseline, which we do in the experimental section.

3 Problem Definition

Preliminaries: 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.

Cascade model: We assume that information on the two campaigns propagates in the network following the independent-cascade model [18]. For instance, consider the first campaign (the process for the second campaign is analogous): we assume that there exists a set of seeds I_1 from which the process begins. Propagation proceeds in rounds. At each round, there exists a set of active vertices A_1 (initially, $A_1 = I_1$), where each vertex $u \in A_1$ attempts to activate each vertex $v \notin A_1$, such that $(u, v) \in E$, with probability $p_1(u, v)$. If the propagation attempt from a vertex u to a vertex v is successful, we say that v propagates the first campaign. At the end of each round, A_1 is set to be the set of vertices that propagated the campaign during the current round.

Given a seed set S, we write $r_1(S)$ and $r_2(S)$ for the vertices that are reached from S using the aforementioned cascade process, for the respective campaign. Note that since this process is random, both $r_1(S)$ and $r_2(S)$ are random variables. Computing the expected number of active vertices is a **#P**-hard problem [8], however, we can approximate it within an arbitrary small factor ϵ, with high probability, via Monte-Carlo simulations. Due to this obstacle, all approximation algorithms that evaluate an objective function over diffusion processes reduce their approximation by an additive ϵ. Throughout this work we avoid repeating this fact for the sake of simplicity of the notation.

Heterogeneous vs. correlated propagations: We also need to specify how the propagation on the two campaigns interact with each other. We consider two settings: In the first setting, we assume that the campaign messages propagate independently of each other. Given an edge $e = (u, v)$, the vertex v is activated on the first campaign with probability $p_1(e)$, given that vertex u is activated on the first campaign. Similarly, v is activated on the second campaign with probability $p_2(e)$, given that u is activated on the second campaign. We refer to this setting as *heterogeneous*.[3] In the second setting we assume that $p_1(e) = p_2(e)$, for each edge e. We further assume that the *coin flips* for the propagation of the two campaigns are totally correlated. Namely, consider an edge $e = (u, v)$, where u is reached by either or both campaigns. Then with probability $p_1(e)$, *any* campaign that has reached u, will also reach v. We refer to this second setting as *correlated*.

Note that in both settings, a vertex may be active by *none*, *either*, or *both* campaigns. This is in contrast to most existing work in competitive viral marketing, where it is assumed that a vertex can be activated by *at most one* campaign. The intuition is that in our setting activation means merely passing a message or posting an article, and it does not imply full commitment to the campaign. We also note that the heterogeneous setting is more *realistic* than the correlated, however, we also study the correlated model as it is mathematically simpler.

Problem definition: We are now ready to state our problem for *balancing information exposure* (BALANCE). Given a directed graph, initial seed sets for both campaigns and a budget, we ask to find additional seeds that would balance the vertices. More formally:

Problem 3.1 (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*

$$\Phi(S_1, S_2) = \mathrm{E}[|V \setminus (r_1(I_1 \cup S_1) \triangle r_2(I_2 \cup S_2))|] .$$

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. Problem 3.1 is defined for both settings, *heterogeneous* and *correlated*. When we need to make explicit the underlying setting we refer to the respective problems by BALANCE-H and BALANCE-C. When referring to BALANCE-H, we denote the objective by Φ_H. Similarly, when referring to BALANCE-C, we write Φ_C. We drop the indices, when we are referring to both models simultaneously.

Computational complexity: As expected, the optimization problem BALANCE turns out to be **NP**-hard for both settings, heterogeneous and correlated. A straightforward way to prove it is by setting $I_2 = V$, so the problems reduce to standard influence maximization. However, we provide a stronger result. Note that instead of maximizing balanced vertices we can equivalently minimize the imbalanced vertices. However, this turns to be a more difficult problem.

Proposition 1. *Assume a graph $G = (V, E, p_1, p_2)$ with two sets I_1 and I_2 and a budget k. It is an **NP**-hard problem to decide whether there are sets S_1 and S_2 such that $|S_1| + |S_2| \leq k$ and $\mathrm{E}[|r_1(I_1 \cup S_1) \triangle r_2(I_2 \cup S_2)|] = 0$.*

This result holds for both models, even when $p_1 = p_2 = 1$. This result implies that the minimization version of the problem is **NP**-hard, and there is no algorithm with multiplicative approximation guarantee. It also implies that BALANCE-H and BALANCE-C are also **NP**-hard. However, we will see later that we can obtain approximation guarantees for these maximization problems.

4 Greedy algorithms yielding approximation guarantees

In this section we propose three greedy algorithms. The first algorithm yields an approximation guarantee of $(1 - 1/e)/2$ for both models. The remaining two algorithms yield a guarantee for the correlated model only.

Decomposing the objective: Recall that the objective function of the BALANCE problem is $\Phi(S_1, S_2)$. In order to show that this function admits an approximation guarantee, we decompose it into two components. To do that, assume that we are given initial seeds I_1 and I_2, and let us write

[3] Although *independent* is probably a better term than *heterogeneous*, we adopt the latter to avoid any confusion with the independent-cascade model.

$X = r_1(I_1) \cup r_2(I_2), Y = V \setminus X$. Here X are vertices reached by any initial seed in the two campaigns and Y are the vertices that are not reached at all. Note that X and Y are random variables. Since X and Y partition V, we can decompose the score $\Phi(S_1, S_2)$ as

$$\begin{aligned}
&\Phi(S_1, S_2) = \Omega(S_1, S_2) + \Psi(S_1, S_2), \quad \text{where} \\
&\Omega(S_1, S_2) = \mathrm{E}\left[\left|X \setminus (r_1(I_1 \cup S_1) \bigtriangleup r_2(I_2 \cup S_2))\right|\right], \\
&\Psi(S_1, S_2) = \mathrm{E}\left[\left|Y \setminus (r_1(I_1 \cup S_1) \bigtriangleup r_2(I_2 \cup S_2))\right|\right].
\end{aligned}$$

We first show that $\Omega(S_1, S_2)$ is monotone and submodular. It is well-known that for maximizing a function that has these two properties under a size constraint, the greedy algorithm computes an $(1 - \frac{1}{e})$ approximate solution [24].

Lemma 2. $\Omega(S_1, S_2)$ *is monotone and submodular.*

We are ready to discuss our algorithms.

Algorithm 0: ignore Ψ. Our first algorithm is very simple: instead of maximizing Φ, we maximize Ω, i.e., we ignore any vertices that are made imbalanced during the process. Since Ω is submodular and monotone we can use the greedy algorithm. If we then compare the obtained result with the empty solution, we get the promised approximation guarantee. We refer to this algorithm as Cover.

Proposition 3. *Let $\langle S_1^*, S_2^* \rangle$ be the optimal solution maximizing Φ. Let $\langle S_1, S_2 \rangle$ be the solution obtained via greedy algorithm maximizing Ω. Then*

$$\max\{\Phi(S_1, S_2), \Phi(\emptyset, \emptyset)\} \ge \frac{1 - 1/e}{2} \Phi(S_1^*, S_2^*).$$

Algorithm 1: force common seeds. Ignoring the Ψ term may prove costly as it is possible to introduce a lot of new imbalanced vertices. The idea behind the second algorithm is to force $\Psi = 0$. We do this by either adding the same seeds to both campaigns, or adding a seed that is covered by an opposing campaign. This algorithm has guarantees only in the correlated setting with even budget k but in practice we can use the algorithm also for the heterogeneous setting. We refer to this algorithm as Common and the pseudo-code is given in Algorithm 1.

Algorithm 1: Common, greedy algorithm that only adds common seeds

```
$S_1 \leftarrow S_2 \leftarrow \emptyset$;
while $|S_1| + |S_2| \le k$ do
    $c \leftarrow \arg\max_c \Phi(S_1 \cup \{c\}, S_2 \cup \{c\})$;
    $s_1 \leftarrow \arg\max_{s \in I_1} \Phi(S_1, S_2 \cup \{s\})$;
    $s_2 \leftarrow \arg\max_{s \in I_2} \Phi(S_1 \cup \{s\}, S_2)$;
    add the best option among $\langle c, c \rangle$, $\langle \emptyset, s_1 \rangle$, $\langle s_2, \emptyset \rangle$ to $\langle S_1, S_2 \rangle$ while respecting the budget.
```

We first show in the following lemma that adding common seeds may halve the score, in the worst case. Then, we use this lemma to prove the approximation guarantee

Lemma 4. *Let $\langle S_1, S_2 \rangle$ be a solution to* BALANCE-C*, with an even budget k. There exists a solution $\langle S_1', S_2' \rangle$ with $S_1' = S_2'$ such that $\Phi_C(S_1', S_2') \ge \Phi_C(S_1, S_2)/2$.*

It is easy to see that the greedy algorithm satisfies the conditions of the following proposition.

Proposition 5. *Assume an iterative algorithm where at each iteration, we add one or two vertices to our solution until our constraints are met. Let S_1^i, S_2^i be the sets after the i-th iteration, $S_1^0 = S_2^0 = \emptyset$. Let $\eta_i = \Phi_C(S_1^i, S_2^i)$ be the cost after the i-th iteration. Assume that $\eta_i \ge \eta_{i-1}$. Assume further that for $i = 1, \ldots, k/2$ it holds that $\eta_i \ge \Phi_C(S_1^{i-1} \cup \{c\}, S_2^{i-1} \cup \{c\})$. Then the algorithm yields $(1 - 1/e)/2$ approximation.*

Algorithm 2: common seeds as baseline. Not allowing new imbalanced vertices may prove to be too restrictive. We can relax this condition by allowing new imbalanced vertices as long as the gain is at least as good as adding a common seed. We refer to this algorithm as Hedge and the pseudo-code is given in Algorithm 2. The approximation guarantee for this algorithm—in the correlated setting and with even budget—follows immediately from Proposition 5 as it also satisfies the conditions.

Algorithm 2: Hedge, greedy algorithm, where each step is as good as adding the best common seed

```
$S_1 \leftarrow S_2 \leftarrow \emptyset$;
while $|S_1| + |S_2| \leq k$ do
    $c \leftarrow \arg\max_c \Phi(S_1 \cup \{c\}, S_2 \cup \{c\})$;
    $s_1 \leftarrow \arg\max_s \Phi(S_1, S_2 \cup \{s\})$;
    $s_2 \leftarrow \arg\max_s \Phi(S_1 \cup \{s\}, S_2)$;
    add the best option among $\langle c, c\rangle$, $\langle \emptyset, s_1\rangle$, $\langle s_2, \emptyset\rangle$, $\langle s_2, s_1\rangle$, to $\langle S_1, S_2\rangle$ while respecting the budget.
```

5 Experimental evaluation

In this section, we evaluate the effectiveness of our algorithms on real-world datasets. We focus on (i) analyzing the quality of the seeds picked by our algorithms in comparison to other heuristic approaches and baselines; (ii) analyzing the efficiency and the scalability of our algorithms; and (iii) providing anecdotal examples of the obtained results. Although we setup our experiments in order to mimic social behavior, we note that fully realistic experiments would entail the ability to intervene in the network, select seeds, and observe the resulting cascades. This, however, is well beyond our capacity and the scope of the paper.

In all experiments we set k to range between 5 and 50 with a step of 5. We report averages over 1 000 random simulations of the cascade process.

Datasets: To evaluate the effectiveness of our algorithms, we run experiments on real-world data collected from twitter. Let $G = (V, E)$ be the twitter follower graph. A directed edge $(u, v) \in E$ indicates that user v follows u; note that the edge direction indicates the "information flow" from a user to their followers. We define a cascade $G_X = (X, E_X)$ as a graph over the set of users $X \subseteq V$ who have retweeted at least one hashtag related to a topic (e.g., US elections). An edge $(u, v) \in E_X \subseteq E$ indicates that v retweeted u.

We use datasets from six topics with opposing viewpoints, covering politics (`US-elections`, `Brexit`, `ObamaCare`), policy (`Abortion`, `Fracking`), and lifestyle (`iPhone`, focusing on iPhone vs. Samsung). All datasets are collected by filtering the twitter streaming API (1% random sample of all tweets) for a set of keywords used in previous work [20]. For each dataset, we identify two sides (indicating the two view-points) on the retweet graph, which has been shown to capture best the two opposing sides of a controversy [12]. Details on the statistics of the dataset can be found at the supplementary material.

After building the graphs, we need to estimate the diffusion probabilities for the heterogeneous and correlated models. Note that the estimation of the diffusion probabilities is orthogonal to our contribution in this paper. For the sake of concreteness we have used the approach described below. One could use a different, more advanced, method; our methods are still applicable.

Let $q_1(v)$ and $q_2(v)$ be an *a priori* probability of a user v retweeting sides 1 and 2, respectively. These are measured from the data by looking at how often a user retweets content from users and keywords that are discriminative of each side. For example, for `US-elections`, the discriminative users and keywords for side Hillary would be @hillaryclinton and #imwither, and for Trump, @realdonaldtrump and #makeamericagreatagain. The probability that user v retweets user u (cascade probability) is then defined as

$$p_i(u, v) = \alpha\, q_i(v) + (1 - \alpha) \left(\frac{R(u, v) + 1}{R(v) + 2} \right), \quad i = 1, 2,$$

where $R(u, v)$ is the number of times v has retweeted u, and $R(v)$ is the total number of retweets of user v. The cascade probabilities p_i capture the fact that users retweet content if they see it from their friends (term $\frac{R(u,v)+1}{R(v)+2}$) or based on their own biases (term $q_i(v)$). The additive terms in the numerator and denominator provide an additive smoothing by Laplace's rule of succession.

We set the value of α to 0.8 for the heterogeneous setting. For $\alpha = 0$ the edge probabilities become equal for the two campaigns, which is our assumption for the correlated setting.

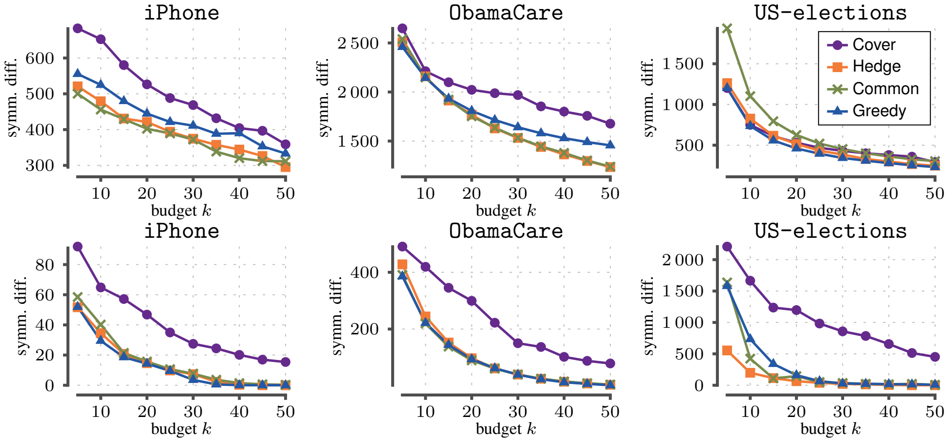


Figure 1: Expected symmetric difference $n - \Phi_C$ as a function of the budget k. Top row, heterogeneous model, bottom row: Correlated model. Low values are better.

Baselines. We use 5 different baselines. The first baseline, BBLO, is an adaptation of the framework by Borodin et al. [5]. This framework requires an objective function as input, and here we use our objective function Φ. The framework works as follows: The two campaigns are given a budget $k/2$ on the number of seeds that they can select. At each round, we select a vertex v for S_1, optimizing $\Phi(S_1 \cup \{v\}, S_2)$, and a vertex w for S_2, optimizing $\Phi(S_1, S_2 \cup \{w\})$. We should stress that the theoretical guarantees by [5] do not apply because our objective is not submodular.

The next two heuristics add a set of common seeds to both campaigns. We run a greedy algorithm for campaign $i = 1, 2$ to select the set S_i' with the $\ell \gg k$ vertices P_i that optimizes the function $r_i(S_i' \cup I_i)$. We consider two heuristics: Union selects S_1 and S_2 to be equal to the $k/2$ first distinct vertices in $S_1' \cup S_2'$ while Intersection selects S_1 and S_2 to be equal to $k/2$ first vertices in $S_1' \cap S_2'$. Here the vertices are ordered based on their discovery time.

Finally, HighDegree selects the vertices with the largest number of followers and assigns them alternately to the two cascades; and Random assigns $k/2$ random seeds to each campaign.

In addition to the baselines, we also consider a simple greedy algorithm Greedy. The difference between Cover and Greedy is that, in each iteration, Cover adds the seed that maximizes Ω, while Greedy adds the seed that maximizes Φ. We can only show an approximation guarantee for Cover but Greedy is a more intuitive approach, and we use it as a heuristic.

Comparison of the algorithms. We start by evaluating the quality of the sets of seeds computed by our algorithms, i.e., the number of equally-informed vertices.

Heterogeneous setting. We consider first the case of heterogeneous networks. The results for the selected datasets are shown in Figure 1. Full results are shown in the supplementary material. Instead of plotting Φ, we plot the number of the remaining unbalanced vertices, $n - \Phi$, as it makes the results easier to distinguish; i.e., an optimal solution achieves the value 0.

The first observation is that the approximation algorithm Cover performs, in general, worse than the other two heuristics. This is due to the fact that Cover does not optimize directly the objective function. Hedge performs better than Greedy, in general, since it examines additional choices to select. The only deviation from this picture is for the US-elections dataset, where the Greedy outperforms Hedge by a small factor. This may due to the fact that while Hedge has more options, it allocates seeds in batches of two.

Correlated setting. Next we consider correlated networks. We experiment with the three approximation algorithms Cover, Common, Hedge, and the heuristic Greedy. The results are shown in Figure 1. Cover performs again the worst since it is the only method that introduces new unbalanced vertices without caring about their cardinality. Its variant, Greedy, performs much better in practice even though it does not provide an approximation guarantee. The algorithms Common, Greedy, and Hedge perform very similar to each other without a clear winner.

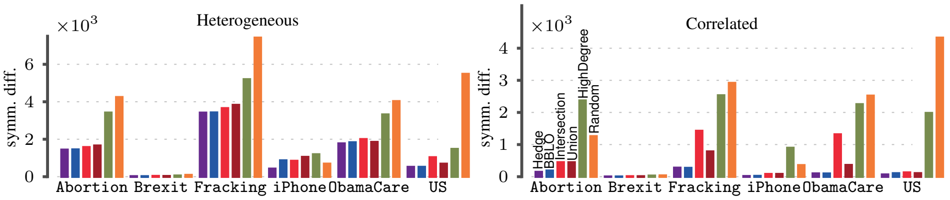


Figure 2: Expected symmetric difference $n - \Phi$ of Hedge and the baselines. $k = 20$. Low values are better.

Comparison with baselines. Our next step is to compare against the baselines. For simplicity, we focus on $k = 20$; the overall conclucions hold for other budgets. The results for Hedge versus the five baselines are shown in Figure 2.

From the results we see that BBLO is the best competitor: its scores are the closest to Hedge, and it receives slightly better scores in 3 out of 12 cases. The competitiveness is not surprising because we specifically set the objective function in BBLO to be $\Phi(S_1, S_2)$. The Intersection and Union also perform well but are always worse than Hedge. Random is unpredictable but always worse than Hedge. In the case of heterogeneous networks, Hedge selects seeds that leave less unbalanced vertices, by a factor of two on average, compared to the seeds selected by the HighDegree method. For correlated networks, our method outperforms the two baselines by an order of magnitude. The actual values of this experiment can be found in the supplementary material.

Running time. We proceed to evaluate the efficiency and the scalability of our algorithms. We observe that all algorithms have comparable running times and good scalability. More information can be found in the supplementary material.

Use case with `Fracking`. We present a qualitative case-study analysis for the seeds selected by our algorithm. We highlight the `Fracking` dataset, even though we applied similar analysis to the other datasets as well (the results are given in the supplementary material of the paper). Recall that for each dataset we identify two sides with opposing views, and a set of initial seeds for each side (I_1 and I_2). We consider the users in the initial seeds I_1 (side supporting fracking), and summarize the text of all their Twitter profile descriptions in a word cloud. The result, contains words that are used to emphasize the benefits of fracking (energy, oil, gas, etc.). We then draw a similar word cloud for the users identified by the Hedge algorithm as seed nodes in the sets S_1 and S_2 ($k = 50$). The result, contains a more balanced set of words, which includes many words used to underline the environmental dangers of fracking. We use word clouds as a qualitative case study to complement our quantitative results and to provide more intuition about our problem statement, rather than an alternative quantitative measure.

6 Conclusion

We presented the first study of the problem of balancing information exposure in social networks using techniques from the area of information diffusion. Our approach has several novel aspects. In particular, we formulate our problem by seeking to optimize a *symmetric difference* function, which is neither monotone nor submodular, and thus, not amenable to existing approaches. Additionally, while previous studies consider a setting with 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.

Our work provides several directions for future work. One interesting problem is to improve the approximation guarantee for the problem we define. Second, we would like to extend the problem definition for more than two campaigns and design approximation algorithms for that case. Finally, we believe that it is worth studying the BALANCE problem under complex diffusion models that capture more realistic social behavior in the presence of multiple campaigns. One such extension is to consider propagation probabilities on the edges that are dependent in the past behavior of the nodes with respect to the two campaigns, e.g., one could consider Hawkes processes [28].

Acknowledgments. This work has been supported by the Academy of Finland projects “Nestor” (286211) and “Agra” (313927), and the EC H2020 RIA project “SoBigData” (654024).

References

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

[2] N. Alon, M. Feldman, A. D. Procaccia, and M. Tennenholtz. A note on competitive diffusion through social networks. *IPL*, 110(6):221–225, 2010.

[3] S. Bharathi, D. Kempe, and M. Salek. Competitive influence maximization in social networks. In *WINE*, 2007.

[4] A. Borodin, Y. Filmus, and J. Oren. Threshold models for competitive influence in social networks. In *WINE*, 2010.

[5] A. Borodin, M. Braverman, B. Lucier, and J. Oren. Strategyproof mechanisms for competitive influence in networks. In *WWW*, pages 141–150, 2013.

[6] C. Budak, D. Agrawal, and A. El Abbadi. Limiting the spread of misinformation in social networks. In *WWW*, pages 665–674, 2011.

[7] T. Carnes, C. Nagarajan, S. M. Wild, and A. Van Zuylen. Maximizing influence in a competitive social network: a follower's perspective. In *EC*, 2007.

[8] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD*, pages 1029–1038, 2010.

[9] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political Polarization on Twitter. In *ICWSM*, 2011.

[10] P. Dubey, R. Garg, and B. De Meyer. Competing for customers in a social network: The quasi-linear case. In *WINE*, 2006.

[11] M. Farajtabar, X. Ye, S. Harati, L. Song, and H. Zha. Multistage campaigning in social networks. In *NIPS*, pages 4718–4726. 2016.

[12] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy in social media. In *WSDM*, pages 33–42, 2016.

[13] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Reducing controversy by connecting oppposing views. In *WSDM*, 2017.

[14] R. K. Garrett. Echo chambers online?: Politically motivated selective exposure among internet news users1. *JCMC*, 14(2):265–285, 2009.

[15] J. Gottfried and E. Shearer. News use across social media platforms 2016. *Pew Research Center*, 2016.

[16] S. Goyal, H. Heidari, and M. Kearns. Competitive contagion in networks. *Games and Economic Behavior*, 2014.

[17] R. Jie, J. Qiao, G. Xu, and Y. Meng. A study on the interaction between two rumors in homogeneous complex networks under symmetric conditions. *Physica A*, 454:129–142, 2016.

[18] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[19] Q. V. Liao and W.-T. Fu. Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions. In *CHI*, pages 2745–2754, 2014.

[20] H. Lu, J. Caverlee, and W. Niu. Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In *CIKM*, pages 213–222, 2015.

[21] W. Lu, W. Chen, and L. V. Lakshmanan. From competition to complementarity: comparative influence diffusion and maximization. *PVLDB*, 9(2):60–71, 2015.

[22] A. Morales, J. Borondo, J. Losada, and R. Benito. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos*, 25(3), 2015.

[23] S. A. Myers and J. Leskovec. Clash of the contagions: Cooperation and competition in information diffusion. In *ICDM*, pages 539–548, 2012.

[24] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions – I. *Mathematical Programming*, 14(1):265–294, 1978.

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

[26] V. Tzoumas, C. Amanatidis, and E. Markakis. A game-theoretic analysis of a competitive diffusion process over social networks. In *WINE*, 2012.

[27] I. Valera and M. Gomez-Rodriguez. Modeling adoption of competing products and conventions in social media. In *ICDM*, 2015.

[28] A. Zarezade, A. Khodadadi, M. Farajtabar, H. R. Rabiee, and H. Zha. Correlated cascades: Compete or cooperate. In *AAAI*, pages 238–244, 2017.

Balancing information exposure in social networks

Supplementary material

Kiran Garimella
Aalto University & HIIT
Helsinki, Finland
kiran.garimella@aalto.fi

Aristides Gionis
Aalto University & HIIT
Helsinki, Finland
aristides.gionis@aalto.fi

Nikos Parotsidis
University of Rome Tor Vergata
Rome, Italy
nikos.parotsidis@uniroma2.it

Nikolaj Tatti
Aalto University & HIIT
Helsinki, Finland
nikolaj.tatti@aalto.fi

A Proof of Proposition 1

Proof. To prove the hardness we will use SET COVER. Here, we are given a universe U and family of sets $C_1, \ldots, C_\ell$, and we are asked to select k sets covering the universe U.

To map this instance to our problem, we first define vertex set V to consist of 3 parts, V_1, V_2 and V_3. The first part corresponds to the universe U. The second part consists of k copies of ℓ vertices, ith vertex in jth copy corresponds to C_i. The third part consists of k vertices b_j. The edges are as follows: a vertex v in the jth copy, corresponding to a set C_i is connected to the vertices corresponding to the elements in C_i, furthermore v is connected to b_j. We set $p_1 = p_2 = 1$. The initial seeds are $I_1 = \emptyset$ and $I_2 = V_1 \cup V_3$. We set the budget to $2k$.

Assume that there is a k-cover, $C_{i_1}, \ldots, C_{i_k}$. We set

$$S_1 = S_2 = \left\{ \text{vertex corresponding to } C_{i_j} \text{ in } j\text{th copy } \right\}.$$

It is easy to see that the imbalanced vertices in I_2 are exposed to the first campaign. Moreover, S_1 and S_2 do not introduce new imbalanced vertices. This makes the objective equals to 0.

Assume that there exists a solution S_1 and S_2 with a zero cost. We claim that $|S_1 \cap (V_1 \cup V_2)| \leq k$. To prove this, first note that $S_1 \cap V_2 = S_2 \cap V_2$, as otherwise vertices in V_2 are left unbalanced. Let $m = |S_1 \cap V_2|$. Since V_3 must be balanced and each vertex in V_2 has only one edge to a vertex in V_3, there at least k vertices in $|S_1 \cap \{V_2 \cup V_3\}|$, that is, we must have $|S_1 \cap V_3| \geq k - m$. Let us write $d_{ij} = |S_i \cap V_j|$. The budget constraints guarantee that

$$d_{11} + d_{12} + d_{22} + d_{13} \leq \sum_{ij} d_{ij} \leq 2k,$$

which can be rewritten as

$$d_{11} + d_{12} \leq 2k - d_{22} - d_{13} \leq 2k - m - (k - m) = k.$$

Construct C as follows: for each $S_1 \cap V_2$, select the set that correponds to the vertex, for each $S_1 \cap V_1$, select any set that contain this vertex (there is always at least one set, otherwise the problem is trivially false). Since V_1 must be balanced, C is a k-cover of U. □

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

B Proof of Lemma 2

Before providing the proof, as a technicality, note that submodularity is usually defined for functions with one argument. Namely, given a universe of items U, we consider functions of the type $f : 2^U \to \mathbb{R}$. However, by taking $U = V \times \{1, 2\}$ we can equivalently write our objectives as functions with one argument, i.e., $\Phi, \Omega, \Psi : 2^U \to \mathbb{R}$.

Proof. The objective counts 3 types of vertices: (*i*) vertices covered by both initial seeds, (*ii*) additional vertices covered by I_1 and S_2, and (*iii*) additional vertices covered by I_2 and S_1. This allows us to decompose the objective as

$$\Omega(S_1, S_2) = \mathrm{E}[|A| + |B| + |C|], \quad \text{where}$$
$$A = r_1(I_1) \cap r_2(I_2), \quad B = (r_1(I_1) \setminus r_2(I_2)) \cap r_2(S_2), \quad C = (r_2(I_2) \setminus r_1(I_1)) \cap r_1(S_1).$$

Note that A does not depend on S_1 and S_2. B grows in size as we add more vertices to S_2, and C grows in size as we add more vertices to S_1. This proves that the objective is monotone.

To prove the submodularity, let us introduce some notation: given a set of edges F, we write $r(S; F)$ to be the set of vertices that can be reached from S via F. This allows us to define

$$\begin{aligned} A(F_1, F_2) &= r(I_1; F_1) \cap r(I_2; F_2), \\ B(F_1, F_2) &= (r(I_1; F_1) \setminus r(I_2; F_2)) \cap r(S_2; F_2), \\ C(F_1, F_2) &= (r(I_2; F_2) \setminus r(I_1; F_1)) \cap r(S_1; F_1). \end{aligned}$$

The score $\Omega(S_1, S_2)$ can be rewritten as

$$\sum_{F_1, F_2} p(F_1, F_2)(|A(F_1, F_2)| + |B(F_1, F_2)| + |C(F_1, F_2)|),$$

where $p(F_1, F_2)$ is the probability of F_1 being the realization of the edges for the first campaign and F_2 being the realization of the edges for the second campaign.

The first term $A(F_1, F_2)$ does not depend on S_1 or S_2. The second term is submodular as a function of S_2 and does not depend of S_1. The third term is submodular as a function of S_1 and does not depend of S_2. Since any linear combination of submodular function weighted by positive coefficients is also submodular, this completes the proof. □

C Proof of Proposition 3

Proof. Write $c = 1 - 1/e$. Let $\langle S_1', S_2' \rangle$ be the optimal solution maximizing Ω. Lemma 2 shows that $\Omega(S_1, S_2) \geq c\Omega(S_1', S_2')$.

Note that $\Psi(\emptyset, \emptyset) \geq \Psi(S_1^*, S_2^*)$ as the first term is the average of vertices not affected by the initial seeds. Thus,

$$\begin{aligned} \Phi(S_1^*, S_2^*) = \Omega(S_1^*, S_2^*) + \Psi(S_1^*, S_2^*) &\leq \Omega(S_1', S_2') + \Psi(S_1^*, S_2^*) \\ &\leq \Omega(S_1', S_2') + \Psi(\emptyset, \emptyset) \leq \Omega(S_1, S_2)/c + \Psi(\emptyset, \emptyset) \\ &\leq \Omega(S_1, S_2)/c + \Psi(\emptyset, \emptyset)/c \\ &\leq (2/c) \max\{\Omega(S_1, S_2), \Psi(\emptyset, \emptyset)\} \\ &\leq (2/c) \max\{\Phi(S_1, S_2), \Phi(\emptyset, \emptyset)\}, \end{aligned}$$

which completes the proof. □

D Proof of Lemma 4

Proof. As we are dealing with the correlated setting, we can write $r(S) = r_1(S) = r_2(S)$. Our first step is to decompose $\omega = \Phi_C(S_1, S_2)$ into several components. To do so, we partition the vertices based on their reachability from the initial seeds,

$$\begin{aligned} A &= r(I_1) \cap r(I_2), & B &= r(I_1) \setminus r(I_2), \\ C &= r(I_2) \setminus r(I_1), & D &= V \setminus (r(I_1) \cup r(I_2)). \end{aligned}$$

Note that these are all random variables. If $S_1 = S_2 = \emptyset$, then $\Phi_C(S_1, S_2) = \mathrm{E}_C[|A| + |D|]$. More generally, S_1 may balance some vertices in C, and S_2 may balance some vertices in B. We may also introduce new imbalanced vertices in D. To take this into account we define

$$\begin{aligned} B' &= B \cap r(S_2), & C' &= C \cap r(S_1), \\ D' &= D \setminus (r(S_1) \bigtriangleup r(S_2)). \end{aligned}$$

We can express the cost of $\Phi_C(S_1, S_2)$ as

$$\omega = \Phi_C(S_1, S_2) = \mathrm{E}_C[|A| + |B'| + |C'| + |D'|]\,.$$

Split $S_1 \cup S_2$ in two equal-size sets, T and Q, and define

$$\omega_1 = \Phi_C(T, T), \quad \omega_2 = \Phi_C(Q, Q).$$

We claim that $\omega \leq \omega_1 + \omega_2$. This proves the proposition, since $\omega_1 + \omega_2 \leq 2 \max\{\omega_1, \omega_2\}$.

To prove the claim let us first split T and Q,

$$T_1 = T \cap S_1,\ T_2 = T \cap S_2,\ Q_1 = Q \cap S_1,\ Q_2 = Q \cap S_2.$$

Our next step is to decompose ω_1 and ω_2, similar to ω. To do that, we define

$$\begin{aligned} B_1 &= B \cap r(T_2), & B_2 &= B \cap r(Q_2), \\ C_1 &= C \cap r(T_1), & C_2 &= C \cap r(Q_1). \end{aligned}$$

Note that, the pair $\langle T, T \rangle$ does not introduce new imbalanced nodes. This leads to

$$\omega_1 = \Phi_C(T, T) = \mathrm{E}_C[|A| + |B_1| + |C_1| + |D|]\,,$$

and similarly,

$$\omega_2 = \Phi_C(Q, Q) = \mathrm{E}_C[|A| + |B_2| + |C_2| + |D|]\,.$$

To prove $\omega \leq \omega_1 + \omega_2$, note that $|D'| \leq |D|$. In addition,

$$\begin{aligned} |B'| &= |B \cap (r(T_2) \cup r(Q_2))| \\ &\leq |B \cap r(T_2)| + |B \cap r(Q_2)| = |B_1| + |B_2| \end{aligned}$$

and

$$\begin{aligned} |C'| &= |C \cap (r(T_1) \cup r(Q_1))| \\ &\leq |C \cap r(T_1)| + |C \cap r(Q_1)| = |C_1| + |C_2|. \end{aligned}$$

Combining these inequalities proves the proposition. □

E Proof of Proposition 5

To prove the proposition, we need the following technical lemma, which is a twist of a standard technique for proving the approximation ratio of the greedy algorithm on submodular functions.

Lemma 1. *Assume a universe U. Let $f : 2^U \to \mathbb{R}$ be a positive function. Let $T \subseteq U$ be a set with k elements. Let $C_0 \subseteq \cdots \subseteq C_k$ be a sequence of subsets of U. Assume that $f(C_i) \geq \max_{t \in T} f(C_{i-1} \cup \{t\})$.*

Assume further that for each $i = 1, \ldots, k$, we can decompose f as $f = g_i + h_i$ such that

1. *g_i is submodular and monotonically increasing function,*
2. *$h_i(W) = h_i(C_{i-1})$, for any $W \subseteq T \cup C_{i-1}$.*

Then $f(C_k) \geq (1 - 1/e) f(T)$.

Proof. The assumptions of the propositions imply

$$\begin{aligned}
f(T) &= g_i(T) + h_i(T) \\
&= g_i(T) + h_i(C_{i-1}) \\
&\leq g_i(C_{i-1}) + h_i(C_{i-1}) + \sum_{t \in T} g_i(C_{i-1} \cup \{t\}) - g_i(C_{i-1}) \\
&= f(C_{i-1}) + \sum_{t \in T} h_i(C_{i-1}) + g_i(C_{i-1} \cup \{t\}) - g_i(C_{i-1}) - h_i(C_{i-1}) \\
&= f(C_{i-1}) + \sum_{t \in T} h_i(C_{i-1} \cup \{t\}) + g_i(C_{i-1} \cup \{t\}) - g_i(C_{i-1}) - h_i(C_{i-1}) \\
&= f(C_{i-1}) + \sum_{t \in T} f(C_{i-1} \cup \{t\}) - f(C_{i-1}) \\
&\leq f(C_{i-1}) + k(f(C_i) - f(C_{i-1})),
\end{aligned}$$

where the first inequality is due to the submodularity of g_i, and is a standard trick to prove the approximation ratio for the greedy algorithm.

We can rewrite the above inequality as

$$kf(T) + (1-k)f(T) = f(T) \leq f(C_{i-1}) + k(f(C_i) - f(C_{i-1})).$$

Rearranging the terms leads to

$$\frac{k-1}{k}(f(C_{i-1}) - f(T)) \leq f(C_i) - f(T) \quad .$$

Applying induction over i, yields

$$f(C_k) - f(T) \geq \left(\frac{k-1}{k}\right)^k (f(C_0) - f(T)) \geq \frac{1}{e}(f(C_0) - f(T)) \geq -f(T)/e,$$

leading to $f(C_k) \geq (1 - 1/e)f(T)$. □

We can now prove the main claim. Note that since we are using the correlated model, we have $r_1 = r_2$. For notational simplicity, we will write $r = r_1 = r_2$.

Proof of Proposition 5. Let OPT be the cost of the optimal solution. Let D be the solution maximizing $\Phi_C(D, D)$ with $|D| \leq k/2$. Lemma 4 guarantees that $OPT/2 \leq \Phi_C(D, D)$.

In order to apply Lemma 6, we first define the universe U as

$$U = \{\langle u, v\rangle \mid u, v \in V\} \cup \{\langle v, \emptyset\rangle \mid v \in V\} \cup \{\langle \emptyset, v\rangle \mid v \in V\}.$$

The sets are defined as

$$C_i = \left\{\langle v, \emptyset\rangle \mid v \in S_1^i\right\} \cup \left\{\langle \emptyset, v\rangle \mid v \in S_2^i\right\}.$$

Given a set $C \subseteq U$, let us define $\pi_1(C) = \{v \mid \langle v, u\rangle \in C, v \neq \emptyset\}$ to be the union of the first entries in C. Similarly, define $\pi_2(C) = \{v \mid \langle u, v\rangle \in C, v \neq \emptyset\}$.

We can now define f as $f(C) = \Phi_C(\pi_1(C), \pi_2(C))$. To decompose f, let us first write

$$X_i = r(I_1 \cup \pi_1(C_{i-1})) \cup r(I_2 \cup \pi_2(C_{i-1})) = r(I_1 \cup S_1^{i-1}) \cup r(I_2 \cup S_2^{i-1}), \quad Y_i = V \setminus X_i.$$

and set

$$\begin{aligned}
g_i(C) &= \mathrm{E}[|X_i \setminus (r(I_1 \cup \pi_1(C)) \triangle r(I_2 \cup \pi_2(C)))|], \\
h_i(C) &= \mathrm{E}[|Y_i \setminus (r(I_1 \cup \pi_1(C)) \triangle r(I_2 \cup \pi_2(C)))|].
\end{aligned}$$

Finally, we set $T = \{\langle d, d\rangle \mid d \in D\}$.

First note that $f = g_i + h_i$ since $X_i \cap Y_i = \emptyset$. The proof of Lemma 2 shows that g_i is monotonically increasing and submodular.

Let $C \subseteq C_{i-1} \cup T$. If there is a vertex v in $r(I_1 \cup \pi_1(C))$ but not in X_i, then this means v was influenced by $d \in D$. Since $d \in \pi_2(C)$, we have $v \in r(I_2 \cup \pi_2(C))$. That is,

$$r(I_1 \cup \pi_1(C)) \setminus X_i = r(I_2 \cup \pi_2(C)) \setminus X_i.$$

Since Y_i and X_i are disjoint, this gives us

$$\begin{aligned} h_i(C) &= \mathrm{E}[|Y_i \setminus (r(I_1 \cup \pi_1(C)) \triangle r(I_2 \cup \pi_2(C)))|] \\ &= \mathrm{E}[|Y_i \setminus ((r(I_1 \cup \pi_1(C)) \setminus X_i) \triangle (r(I_2 \cup \pi_2(C)) \setminus X_i))|] \\ &= \mathrm{E}[|Y_i|]\,. \end{aligned}$$

That is, $h_i(C)$ is constant for any $C \subseteq C_{i-1} \cup T$. Thus, $h_i(C) = h_i(C_{i-1})$.

Finally, the assumption of the proposition guarantees that $f(C_i) \geq f(C_{i-1} \cup \{t\})$, for $t \in T$.

Thus, these definitions meet all the prerequisites of Lemma 6, guaranteeing that

$$(1 - 1/e)\Phi_C(D, D) \leq \Phi_C(S_1^{k/2}, S_2^{k/2}) \leq \Phi_C(S_1^k, S_2^k).$$

Since $OPT/2 \leq \Phi_C(D, D)$, the result follows. □

F Additional tables and figures related to the experimental evaluation

Table 1: Dataset descriptions, as well as tags and rewteets that were used to collect the data.

USelections: Tweets containing hashtags and keywords identifying the USElections, such as #uselections, #trump2016, #hillary2016, etc. Collected using Twitter 1% sample for 2 weeks in September 2016	
Pro-Hillary	*Pro-Trump*
RT @hillaryclinton, #hillary2016, #clintonkaine2016, #imwithher	RT @realdonaldtrump, #makeamericagreatagain, #trumppence16, #trump2016
Brexit: Tweets containing hashtags #brexit, #voteremain, #voteleave, #eureferendum for all of June 2016, from the 1% Twitter sample.	
Pro-Remain	*Pro-Leave*
#voteremain, #strongerin, #remain, #remaineu, #votein	#voteleave, #strongerout, #leaveeu, #takecontrol, #leave, #voteout
Abortion: Tweets containing hashtags #abortion, #prolife, #prochoice, #anti-abortion, #pro-abortion, #plannedparenthood from Oct 2011 to Aug 2016.	
Pro-Choice	*Pro-Life*
RT @thinkprogress, RT @komenforthecure, RT @mentalabortions, #waronwomen, #nbprochoice, #prochoice, #standwithpp, #reprorights	RT @stevenertelt, RT @lifenewshq, #praytoendabortion, #prolifeyouth, #prolife, #defundplannedparenthood, #defundpp, #unbornlivesmatter
Obamacare: Tweets containing hashtags #obamacare, and #aca from Oct 2011 to Aug 2016.	
Pro-Obamacare	*Anti-Obamacare*
RT @barackobama, RT @lolgop, RT @charlespgarcia, RT @defendobamacare, RT @thinkprogress, #obamacares, #enoughalready, #uniteblue	RT @sentedcruz, RT @realdonaldtrump, RT @mittromney, RT @breitbartnews, RT @tedcruz, #defundobamacare, #makedclisten, #fullrepeal, #dontfundit
Fracking: Tweets containing hashtags and keywords #fracking, 'hydraulic fracturing', 'shale', 'horizontal drilling', from Oct 2011 to Aug 2016.	
Pro-Fracking	*Anti-Fracking*
RT @shalemarkets, RT @energyindepth, RT @shalefacts, #fracknation, #frackingez, #oilandgas, #greatgasgala, #shalegas	RT @greenpeaceuk, RT @greenpeace, RT @ecowatch, #environment, #banfracking, #keepitintheground, #dontfrack, #globalfrackdown, #stopthefrackattack
iPhone vs. Samsung: Tweets containing hashtags #iphone, and #samsung from April (release of Samsung Galaxy S7), and September 2015 (release of iPhone 7).	
Pro-iPhone	*Pro-Samsung*
#iphone	#samsung

Table 2: Dataset statistics. The column $|C|$ refers to the average number of edges in a randomly generated cascade in the correlated case, while $|C_1|$ and $|C_2|$ refer to average number of edges generated in a cascade of the campaigns 1 and 2, respectively, in the heterogeneous case.

Dataset	# Nodes	# Edges	$\|C\|$	$\|C_1\|$	$\|C_2\|$
Abortion	279 505	671 144	2 105	326	1 801
Brexit	22 745	48 830	476	113	390
Fracking	374 403	1 377 085	4 156	1 595	3 103
iPhone	36 742	49 248	4 776	339	4 478
ObamaCare	334 617	1 511 670	6 614	2 404	4 527
US-elections	80 544	921 368	4 697	3 097	12 044

Table 3: Detailed values of the data presented in Figure 2. The data correspond to the absolute value expected symmetric difference $n - \Phi$ of Hedge and the baselines for $k = 20$ across all datasets. Low values are better.

			Heterogeneous setting			
Dataset	Hedge	BBLO	Inters.	Union	HighDeg.	Random
Abortion	1436.090	1447.710	1571.180	1655.580	3414.310	4253.220
Brexit	17.907	17.765	31.850	27.770	54.131	87.341
Fracking	3411.810	3420.700	3651.230	3825.360	5197.060	7449.350
iPhone	421.411	865.126	839.119	1048.090	1189.650	631.543
ObamaCare	1768.560	1828.900	1998.250	1846.750	3315.570	4032.140
US-elections	515.347	516.587	1030.640	685.089	1474.330	5988.160
			Homogeneous setting			
Dataset	Hedge	BBLO	Inters.	Union	HighDeg.	Random
Abortion	144.898	185.569	446.462	444.766	2368.610	1279.100
Brexit	1.232	1.615	9.643	9.374	28.850	34.283
Fracking	275.143	269.404	1423.870	781.994	2529.570	2960.720
iPhone	14.624	19.893	79.854	80.279	895.353	759.629
ObamaCare	97.319	95.062	1314.830	360.103	2253.050	2484.330
US-elections	64.870	103.318	128.586	104.911	1979.79	5325.130

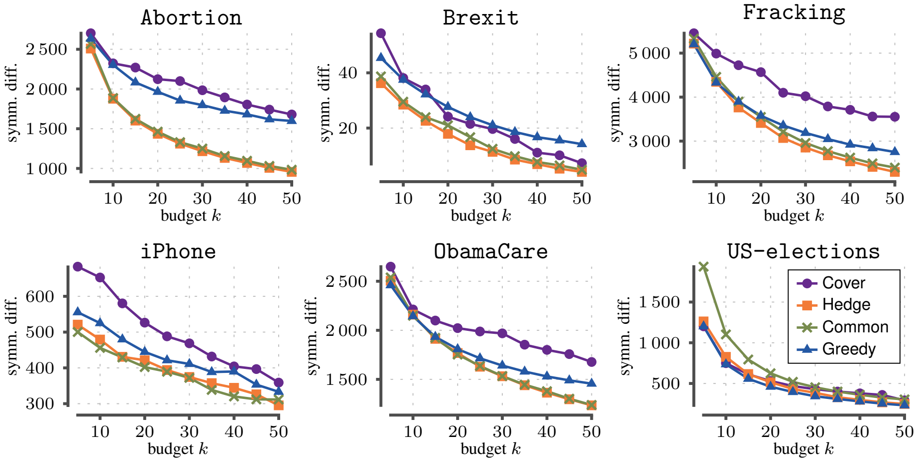


Figure 1: Expected symmetric difference $n - \Phi_H$ as a function of the budget k. Heterogeneous model. Low values are better.

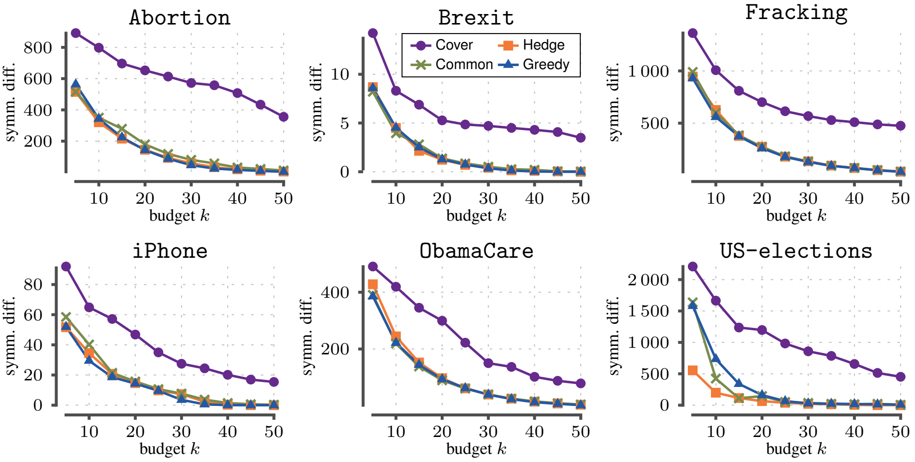


Figure 2: Expected symmetric difference $n - \Phi_C$ as a function of the budget k. Correlated model. Low values are better.

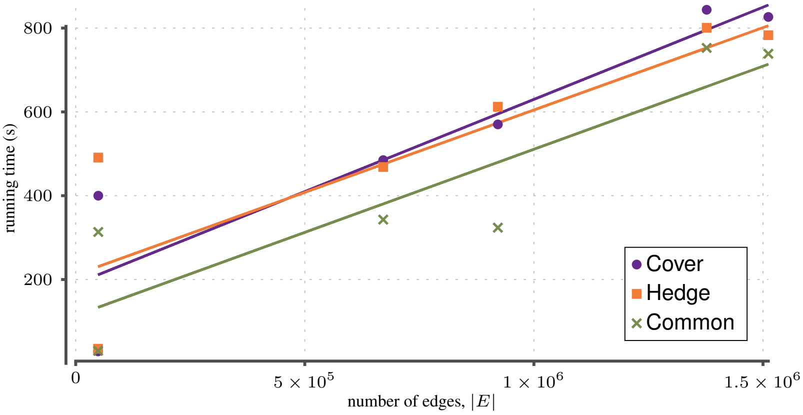


Figure 3: Running time as a function of number of edges. Correlated model with $k = 20$.

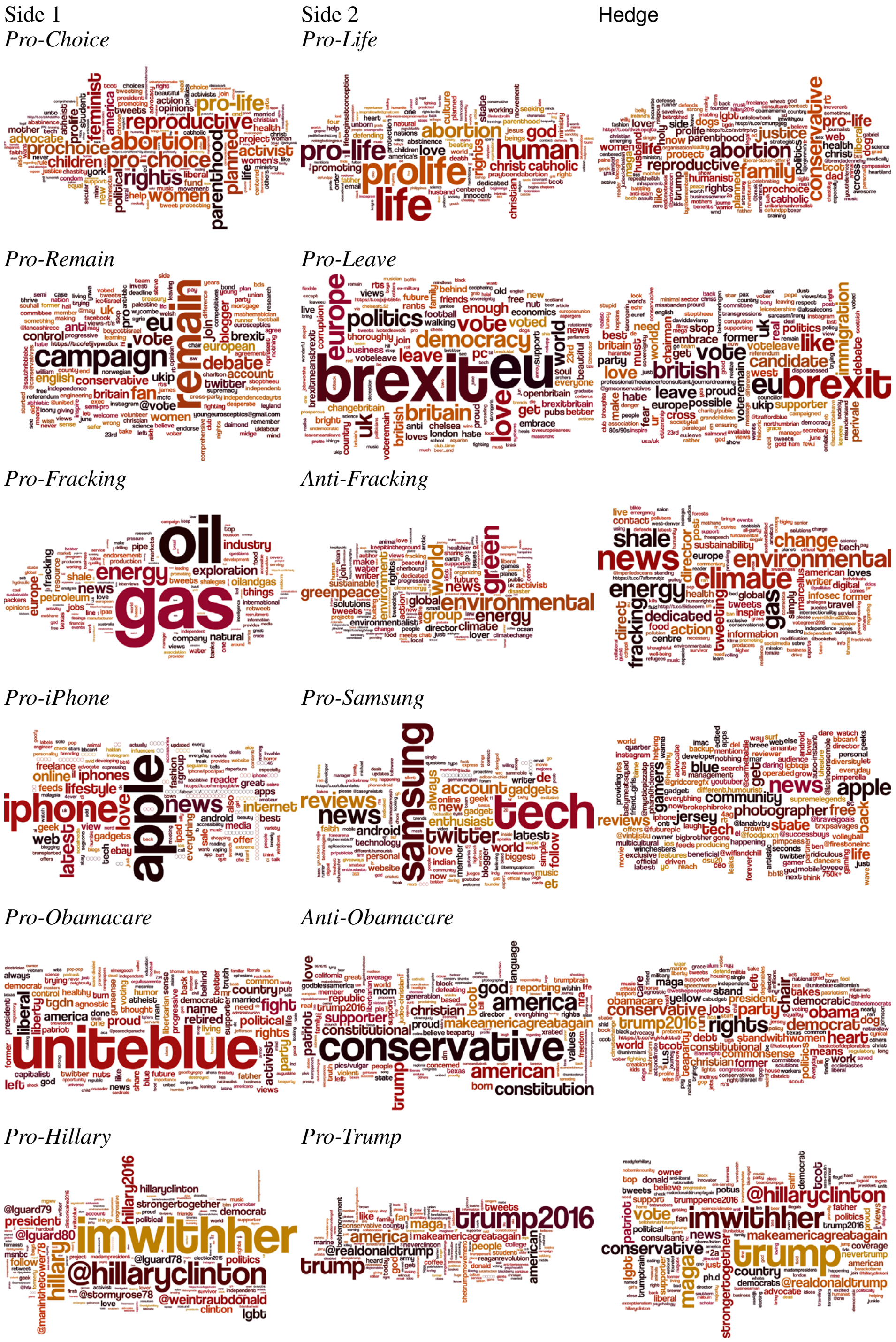


Figure 4: Word clouds of the profiles for the initial seeds, and profiles selected by Hedge.

Publication 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.

Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship

Kiran Garimella
Aalto University
Espoo, Finland
kiran.garimella@aalto.fi

Gianmarco De Francisci Morales
Qatar Computing Research Institute
Doha, Qatar
gdfm@acm.org

Aristides Gionis
Aalto University
Espoo, Finland
aristides.gionis@aalto.fi

Michael Mathioudakis
CNRS LIRIS & INSA Lyon
Lyon, France
michael.mathioudakis@liris.cnrs.fr

ABSTRACT

Echo chambers, i.e., situations where one is exposed only to opinions that agree with their own, are an increasing concern for the political discourse in many democratic countries. This paper studies the phenomenon of political echo chambers on social media. We identify the two components in the phenomenon: the opinion that is shared ("echo"), and the place that allows its exposure ("chamber" — the social network), and examine closely at how these two components interact. We define a production and consumption measure for social-media users, which captures the political leaning of the content shared and received by them. By comparing the two, we find that Twitter users are, to a large degree, exposed to political opinions that agree with their own. We also find that users who try to bridge the echo chambers, by sharing content with diverse leaning, have to pay a "price of bipartisanship" in terms of their network centrality and content appreciation. In addition, we study the role of "gatekeepers," users who consume content with diverse leaning but produce partisan content (with a single-sided leaning), in the formation of echo chambers. Finally, we apply these findings to the task of predicting partisans and gatekeepers from social and content features. While partisan users turn out relatively easy to identify, gatekeepers prove to be more challenging.

ACM Reference format:
Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. In *Proceedings of WWW '18, Lyon, France, April 23–27, 2018,* 10 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Echo chambers have recently emerged as an issue of concern in the political discourse of democratic countries. There is growing concern that, as citizens become more polarized about political issues, they do not hear the arguments of the opposite side, but are rather surrounded by people and news sources who express only opinions they agree with. It is telling that Facebook and ex-U.S. Presidents have recently voiced such concerns.[1] If echo chambers exist, then scholars agree that they are a threat to deliberative democracy, as they cause a disconnect between facts and how they are perceived [34].

In this paper, we study the degree to which echo chambers exist in political discourse on Twitter, and how they are structured. We approach the study in terms of two components: the opinion that is shared on the platform ("echo"), and the place that allows its exposure ("chamber"). The opinion corresponds to *content* items shared by users, while the underlying social *network* is what allows their propagation. We say that an echo chamber exists if *the political leaning of the content that users receive from the network agrees with that of the content they share.*

As there is no consensus on a formal definition in the literature, we opt for this definition, which is general enough and reasonably captures the essence of the phenomenon. There are, however, a few previous works that have studied echo chambers under different perspectives. For instance, previous works have focused either on the differences between the content shared and read by partisans of different sides [3, 18, 19, 33]; the social network structure [21]; or the structure of user interactions, such as blog linking [1] and retweets [10, 15]. We adopt a broader definition in terms of the notion of *content* it is based on (it considers all content shared and produced, not only content pertaining to specific types of interactions, e.g., retweets) and is defined jointly on *content* and *network*.

Specifically, we define production and consumption measures for social media users based on the political leaning of the content *shared with* and *received from* their network. We apply them to several datasets from Twitter, including a large one consisting of over 2.5 billion tweets, which captures 8 years worth of exchanges between politically-savvy users. Our findings indicate there is large correlation between the leaning of content produced and consumed: *echo chambers are prevalent on Twitter.*

WWW '18, Lyon, France
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

[1] E.g., Obama foundation's attempt to address the issue of echo chambers. https://www.engadget.com/2017/07/05/obama-foundation-social-media-echo-chambers

We then proceed to analyze *partisan* users, who produce content with predominantly one-sided leaning,[2] and *bipartisan* users, which instead produce content with both leanings. Our analysis indicates that partisan users enjoy a higher "appreciation" as measured by both network and content features. This finding hints at the existence of a *"price of bipartisanship,"* required to be paid by users who try to bridge echo chambers.

Moreover, we take a closer look on *gatekeeper* users, who consume content of both leanings, but produce content of a single-sided leaning. These users are *border spanners* in terms of location in the social network, who remain aware of the positions of both sides, but align their content with one side. They are a small group, which enjoy higher than average network centrality, while not being very embedded in their community.

Finally, we use these findings for predicting *partisan* and *gatekeeper* users by using features from the content they produce and from their social network. While partisan users are relatively easy to identify, gatekeepers prove to be more challenging.

Our study opens the road for further investigation of the echo chamber phenomenon. While establishing the existence of political echo chambers on Twitter, based on a broad definition and measurements over a large volume of data, it also invites a more nuanced analysis of such phenomenon – one that, instead of categorizing users in terms of partisanship, takes into account a variety of user attitudes (e.g., partisans, gatekeepers, and bipartisans). Such analysis might be crucial to understand how to nudge users towards consuming content that challenges their opinion and thus bridge echo chambers. Furthermore, our study shows the interdependence between content production & consumption and network properties in the context of echo chambers. This finding could help us in revisiting existing models for the dynamics of opinion formation and polarization on social networks [11, 32] that take into account not only the opinion (content) spread over the social network, but also its impact of structure of the network itself.

2 RELATED WORK

Echo chambers. Echo chambers refers to situations where people "hear their own voice" — or, particularly 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 [19, 35], forums [13], and social-media sites [7, 21, 33].

Previous studies have tried to quantify the extent to which echo chambers exist online. For example, in the context of blogs, Gilbert et al. [19] study the comments on a set of political blogs and find that comments disproportionately agree with the author of the blog post. Similar findings were reported by Lawrence et al. [24], who found that partisan bloggers engage with blogs of a narrow spectrum of political views, which agreed with their own. In the context of Twitter, An et al. [2] analyzed the activity of users who engage with political news, and found that "90% of the users [directly follow] news media of only one political leaning", while "their friends' retweets lead them to diversify their news consumption".

In the context of facebook, Bakshy et al. [4] measure the degree to which users with declared political affiliations consume cross-cutting content, i.e., content predominantly posted by users of opposing political affiliation. Content consumption is studied at three levels: (*i*) *potential exposure*, which includes all content shared by the friends of a user; (*ii*) *exposure*, which includes all content appearing in the feed of a a user; and (*iii*) *engagement*, which includes all content that a user clicks. The study finds that, even though users are exposed to a significant amount of cross-cutting content, users opt to engage with less cross-cutting content, a behavior compatible with the theory of *biased assimilation* [26].

However, there is no consistent definition of what an echo chamber represents in the literature. The studies presented above measure different aspects of an echo chamber, and focus *either* on the "echo" (content) [4, 19, 24] *or* the "chamber" (network) [1, 2].

In this paper, we propose measures to identify the existence of an echo chamber by using *both* the content being read/shared *and* the network that enables the content to propagate. Unlike many previous works that focus on measuring only content consumption to quantify the echo-chamber effects, we study content consumption and production jointly at the level of individual users, and examine how different content profiles correlate with the network position of users. Though we are not the first to show the existence of echo chambers on Twitter, to the best of our knowledge, this is the first study to jointly use content and network to characterize echo chambers.

Psychological and algorithmic mechanisms. *Selective exposure theory* [14] — which proposes the concepts of *selective exposure*, *selective perception*, and *selective retention* — is the tendency of individuals to favor information that aligns with their pre-existing views while avoiding contradictory information. *Biased assimilation* [26], 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. All these psychological mechanisms, together with other biases, such as, *algorithmic filtering* and *personalization* [9], are connected to the phenomenon of echo chambers. Understanding how all these phenomena interact with each other and the precise causality relations is beyond the scope of this paper.

Relationship between node and network properties. One of our objectives is to understand the relationship between node properties (user consumption and production) and network properties (e.g., PageRank and clustering coefficient).

Homophily is a central notion in the study of social networks. Given a network and a node feature, homophily refers to the phenomenon where neighboring nodes in the network tend to present similar values of the given feature. Several studies have provided evidence of homophily in social networks [29]. For example, in the context of Twitter, clusters in retweet networks have been found to correlate with the political ideologies of Twitter users [7, 10, 15].

Price of bipartisanship. Hetherington [22] argues that political parties have increased their prominence in the masses by being more partisan. Prior [31] analyzes the role of partisan media to answer the question: "has partisan media created political polarization and led the American public to support more partisan policies

[2] We use "leaning" as a score that quantifies alignment with one political side. Similar terms in the literature include "ideology," "polarity," or "ideological stance."

and candidates?" They find no evidence to support that claim. Conversely, DellaVigna and Kaplan [12] 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.

In this paper, we study the price of being bipartisan, for the first time on social networks. We show that producing content that expresses opinions aligned with both sides of the political divide, has a cost in terms of centrality in the network and content-engagement rate.

Gatekeeping. Gatekeeping is a term commonly used in communication studies to refer to news media sources that act as filters of information [25]. Barzilai-Nahon [8] propose a model based on network theory for gatekeeping which generalizes the concept of gatekeeping for the Internet and applies to all information types (not just news). Several studies have looked at gatekeeping practices on Twitter [23, 37] and conclude that unlike in traditional media, any common user can become a gatekeeper on social media. The definition of these gatekeepers on social media also differs from the traditional gatekeepers in media organizations, due to the alternatives available to social media users.

In our case, we define gatekeepers as users who receive content from both political leanings, but only produce content from a single leaning, thus "filtering" information from one side. To the best of our knowledge, this is the first paper to study the role of gatekeepers of information within echo chambers.

3 DATA

We use a collection of ten different datasets from Twitter, each of which contains a set of tweets on a given topic of discussion. The datasets span over a long period of time and cover a wide range of users and topics, described below. The collection is partitioned into two groups, Political and Non-Political, depending on whether the topic of discussion is politically contentious or not. Moreover, in addition to tweets, for each dataset, we build a network that represents the social connections among users. The size of each dataset in terms of number of tweets and number of distinct users is shown in Table 1. For all the datasets, we perform simple checks to remove bots, using minimum and maximum thresholds for number of tweets per day, followers, friends, and ensure that the account is at least one year old at the time of data collection. More details about the datasets are given below.

Political. Five of the ten Twitter datasets are relevant to well-known political controversies. Three of these datasets, namely `guncontrol`, `obamacare`, and `abortion`, discuss a specific topic. Each dataset is built by collecting tweets posted during specific events that lead to an increased interest in these topics (see Table 1). Using the Archive Twitter Stream grab,[3] we select tweets that contain keywords pertaining to each topic that were posted in a time period of one week around the event (3 days before and 3 days after the event).[4] To focus on users who are actively engaged in the discussion of each topic, we identify the subset of users who have at least 5 tweets about the topic during this time window. We collect all the tweets posted by these users via Twitter's REST API.[5]

Table 1: Description of the datasets.

Topic	#Tweets	#Users	Event
`guncontrol`	19M	7506	Democrat filibuster for gun-control reforms (June 12–18, 2016)[6]
`obamacare`	39M	8773	Obamacare subsidies preserved in U.S. supreme court ruling (June 22–29, 2015)[7]
`abortion`	34M	3995	Supreme court strikes down Texas abortion restrictions (June 27–July 3, 2016)[8]
`combined`	19M	6391	2016 US election result night (Nov 6–12, 2016)
`large`	2.6B	676 996	Tweets from users retweeting a U.S. presidential/vice presidential candidate (from [17], 2009–2016)
`ff`	4M	3204	filtering for these hashtags
`gameofthrones`	5M	2159	
`love`	3M	2940	
`tbt`	28M	12 778	
`foodporn`	8M	3904	

A fourth dataset, named `combined`, is collected in a similar fashion, except that it contains tweets of users who were active during the U.S. presidential election results of 2016 (November 6–12, 2016), and who tweeted at least 5 times about any of the three controversial topics `guncontrol`, `obamacare`, and `abortion`. We also collect all tweets of these users via Twitter's REST API.

Finally, the fifth dataset, named `large`, is a large dataset containing over 2.5 billion tweets from politically active users spanning a period of almost 8 years (2009-2016). Specifically, the dataset consists of all tweets generated by users who retweeted a presidential or vice-presidential candidate from 2008-2016 in the U.S. at least 5 times. The dataset has been used in previous work [17]; we refer to the original paper for more details.

Non-Political. To have a baseline for our measurements over the Political datasets, we also use five datasets that correspond to non-political topics, in particular: `tbt` ("throwback Thursday"), `ff` ("follow Friday"), `gameofthrones`, `love`, and `foodporn`. Each of these topics is associated with a particular hashtag (e.g., #tbt for `tbt`). The datasets are built as follows. First, we parse the tweets in the Internet Archive collection and select tweets that contain the corresponding hashtag for each topic during the month of June 2016. Second, we filter out users who have less than 5 tweets. Third, we obtain all tweets generated by these users. The resulting set of tweets for each topic constitutes one dataset.

Network. For each dataset, we build the directed "follow" graph among users: an edge $(u \rightarrow v)$ indicates that user u follows user v.

Political leaning scores (source polarity). Our analysis relies on characterizing the political leaning of the content consumed and produced by each user. Obtaining a characterization of political leaning for short text snippets, such as tweets, is a very challenging

[3] https://archive.org/details/twitterstream
[4] We use the keyword lists proposed by Lu et al. [27].
[5] https://developer.twitter.com/en/docs/tweets/timelines/overview
[6] https://en.wikipedia.org/wiki/Chris_Murphy_gun_control_filibuster
[7] http://www.bbc.com/news/world-us-canada-33269991
[8] https://www.nytimes.com/2016/06/28/us/supreme-court-texas-abortion.html

problem, in general. To confront this challenge, we use a ground truth of political leaning scores of various news organizations with a presence on social media obtained from Bakshy et al. [4]. Specifically, the data contains a score of political leaning for 500 news domains (e.g., *nytimes.com*) that are most shared on Facebook. The score takes values between 0 and 1 and expresses the fraction of Facebook users who visit these pages that identify themselves as conservative on their Facebook profile. A value close to 1 (0) indicates that the domain has a conservative (liberal) bent in their coverage. For a detailed description of the dataset, we refer the reader to the original publication [4]. We remove a small number of domains that are not owned by news organizations (e.g., *wikipedia.org* or *reddit.com*), and add shortened versions of news domains to the list (e.g. *fxn.ws* for *foxnews.com*). The distribution of source polarity for the 500 domains is shown in Figure 2.

4 MEASURES

This section describes the measures used in our analysis. These measures aim to capture user activity from two perspectives: (*i*) the *content* produced and consumed by a user, and (*ii*) the *network* position of a user, including their interactions with others.

4.1 Content

Content is central in measuring echo chamber effects. In a setting where opinions are polarized between two perspectives – in our case "liberal" and "conservative" – we say that *an echo chamber exists to the degree that users consume content that agrees with their point of view.* To make this definition actionable and quantify the echo chamber effect, we need to model the *political leaning* of content *produced* and *consumed* by users.

For the *content production* of a user u, we consider tweets posted by user u. For the *content consumption* of a user u we consider tweets posted by users whom u follows.

To quantify the political leaning of content posted on Twitter, we consider only messages that contain a link to an online news organization with a known and independently derived political leaning. In particular, we use the dataset of the political leaning scores of news organizations described in Section 3. Based on those scores, we define a polarity score for the content produced and consumed by a user.

Production polarity. For each user u in a given dataset, we consider the set of tweets P_u posted by u that contain links to news organizations of known *political leaning* l_n. We then associate each tweet $t \in P_u$ with leaning $\ell(t) = l_n$. The *production polarity* $p(u)$ of user u is then defined as the average *political leaning* over P_u, i.e.,

$$p(u) = \frac{\sum_{t \in P_u} \ell(t)}{|P_u|}. \tag{1}$$

The value of *production polarity* ranges between 0 and 1. For users who regularly share content from liberal sources, *production polarity* is closer to 0, while for the ones who share content from conservative sources it is closer to 1.

We wish to quantify the extent to which users produce one-sided content. We say that a user is δ-**partisan**, for some value $0 \leq \delta \leq \frac{1}{2}$, if their *production polarity* is within δ from either extreme value

$$\min\{p(u), 1 - p(u)\} \leq \delta. \tag{2}$$

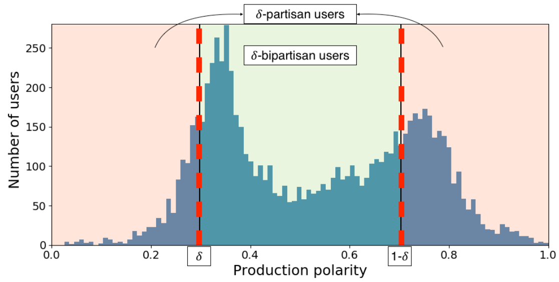


Figure 1: Example showing the definition of δ-partisan users. The dotted red lines are drawn at δ and 1-δ. Users on the left of the leftmost dashed red line or right of the rightmost one are δ-partisan.

The smaller the value of δ the more partisan a user is. Note also that if a user u is δ-partisan then u is also δ'-partisan for $\delta < \delta' \leq \frac{1}{2}$. Users who are not δ-partisan are called δ-bipartisan. Intuitively, δ-partisan users produce content only from one extreme end of the political spectrum, where as δ-bipartisan ones do not. Figure 1 shows an illustration of δ-partisan and δ-bipartisan users.

Production variance. Besides the average *political leaning* of produced tweets, we also measure the *variance* in *political leaning* over the same set of tweets. The objective is to quantify the range of opinions of a user covered by the produced content.

Consumption polarity. Similarly to *production polarity*, we define *consumption polarity* based on the set of tweets $C(u)$ that a user receives on their feed from users they follow. We again focus on tweets that contain a link to a news article from a domain with known source polarity. The *consumption polarity* $c(u)$ of user u is defined as the average *political leaning* of received tweets $C(u)$.

$$c(u) = \frac{\sum_{t \in C_u} \ell(t)}{|C_u|} \tag{3}$$

Values close to 0 indicate consumption of liberal content, while values close to 1 indicate consumption of conservative content. Though *consumption polarity* is defined just by looking at the source polarity of tweets, it also takes the network structure into account and forms the basis for the understanding of the interaction between content and network.

To quantify the extent to which users consume one-sided content, we say that a user is δ-**consumer**, for some value $0 \leq \delta \leq \frac{1}{2}$, if their *consumption polarity* is within δ from either extreme value

$$\min\{c(u), 1 - c(u)\} \leq \delta. \tag{4}$$

The values of *consumption polarity* behave similarly to those of *production polarity*.

Consumption variance. Besides the average *political leaning* of consumed tweets, we also measure the *variance* in *political leaning* over the same set of tweets. The objective is to quantify the range of opinions of a user covered by the consumed content.

Gatekeepers. Gatekeepers are defined in media and communication studies as media sources that act as filters (or 'gatekeepers') of information [25]. In our case, we consider consumption and production of content jointly, and define gatekeepers as users who consume content from both sides of the political spectrum but only produce content from one side. These users block or filter information from one side, and hence can be considered gatekeepers.

Formally, we say that a user u is δ-**gatekeeper** if u is δ-partisan but **not** δ-consumer, i.e.,

$$\min\{p(u), 1 - p(u)\} \leq \delta, \quad \text{and} \quad \min\{c(u), 1 - c(u)\} > \delta. \qquad (5)$$

4.2 Network

Our goal is to understand the interplay of content consumption and production with the position of the users in the network and the global network structure. Thus, to add to the above measures defined using content, we define measures that capture the position of the user in a network and their interactions with other users. We consider the following network measures.

User polarity. We adopt the latent space model proposed by Barberá et al. [7] to estimate a *user polarity* score. This score is based on the assumption that Twitter users prefer to follow politicians whose position on the latent ideological dimension is similar to theirs. For the list of politicians and details on estimating the polarity, please refer to the original paper [7]. Negative (positive) values of the user polarity scores indicate a democrat (republican) leaning and the absolute value of the polarity indicates the degree of support to the respective party.

Network centrality. We employ the well-known PageRank measure [30] to characterize the centrality of a node in a network. PageRank reflects the importance of a node in the follow network, and a higher PageRank can be interpreted as a higher chance of the user to spread its content to its community.

Clustering coefficient. In an undirected graph, the clustering coefficient $cc(u)$ of a node u is defined as the fraction of closed triangles in its immediate neighborhood. Specifically, let d be the degree of node u, and T be the number of closed triangles involving u and two of its neighbors. The clustering coefficient is then defined as $cc(u) = \frac{2T}{d(d-1)}$. Note that, as the networks in our datasets are directed graphs, we consider their undirected version to compute clustering coefficients. A high clustering coefficient for a node indicates that the ego network of the corresponding user is tightly knit, i.e., the node is embedded in a well-connected community.

Retweet/Favorite rate. For a given dataset, the *retweet rate* (*favorite rate*) of a user is the fraction of the tweets of that user that have received at least one retweet (favorite).

Retweet/Favorite volume. For a given dataset, the *retweet volume* (*favorite volume*) of a user is defined as the median number of retweets (favorites) received by their tweets. This is different from the retweet/favorite rate because it indicates the popularity of the content, where as the retweet/favorite rate captures "acceptance" of the user's content.

5 ANALYSIS

In this section, we analyze the datasets described in Section 3 using the measures defined in Section 4 to answer the following questions:

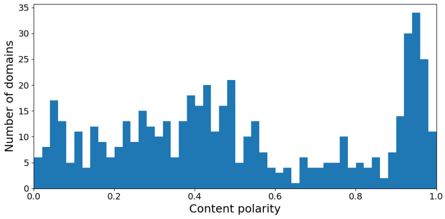


Figure 2: Distribution of source polarity for the 500 news sources considered in the current work [4].

(1) Are there echo chambers, or, are users exposed to content that carries opposite leaning? We answer these questions by looking at the joint distribution of production and consumption polarities (§ 5.1).

(2) Is there an advantage in being partisan? We quantify advantage in terms of network centrality and connectivity (PageRank and clustering coefficient, respectively), as well as in terms of content appreciation (number of retweets and favorited tweets) (§ 5.2).

(3) Who are the users who act as gatekeepers of information in the network? We explore features of these users and examine how they differ from other users. (§ 5.3).

(4) Can we predict if a user is a partisan or a gatekeeper, just by looking at their tweets? We build a classification model that predicts if a user is a partisan or a gatekeeper, leveraging features extracted from the above analysis (§ 5.4).

5.1 Echo chambers: content production and consumption

As discussed in Section 4, the political leaning of produced and consumed content is measured based on the leaning of cited news sources. The distribution of source polarity scores for the news sources is shown in Figure 2. The distribution shows that there are many conservative outlets, and a sizeable number of neutral and liberal outlets.

To explore the values of production and consumption polarities across the datasets, let us examine Figure 3. The top row shows five plots for the POLITICAL datasets, and the bottom row for the NON-POLITICAL ones. Each plot contains three subplots: a two-dimensional scatter-plot in the center and two one-dimensional subplots along the two axes of the scatter-plot.

The distribution of production and consumption polarities of users in the various datasets is shown in the scatter plots of Figure 3. Each point in the scatter-plot corresponds to a user. Recall that lower polarities indicate liberal users, and higher polarities indicate conservative alignment. The color of each point indicates the sign of the user polarity score, as defined by Barberá [6] and described in Section 4 (grey = negative = democrat, yellow = positive = republican). The difference between the two groups of datasets is stark: production and consumption polarities are highly correlated for POLITICAL datasets, which means that users indeed tend to consume content with political leaning aligned to their own. The same does not hold for the Non-Political group, where the correlations are low to non-existent.

How do the production and consumption polarities align with user polarity scores? To explore this, let us turn to the one-dimensional subplots that accompany each scatter-plot. The subplot along the x-axis (y-axis) shows the distributions for production (consumption) polarity for democrats and republicans — as before, defined in terms of the sign of user polarity [6]. We observe that the production and consumption polarities for the Political datasets exhibit clearly separated and bi-modal distributions, while the distributions very much coincide for the Non-Political datasets. This kind of bimodal distribution is also indicative of a divide in the leaning of the content produced and consumed.

Furthermore, let us note that, when the distributions of production and consumption polarities are compared with the source polarity scores in Figure 2, they appear quite different. The production/consumption polarities are more concentrated towards the middle of the spectrum (i.e., there are few very extreme users), and the modes themselves are relatively far from the extremes. In addition, the concentration of the distributions show a preference for one leaning when compared to the distribution of source polarities. This preference can be attributed to personal choice of the user (for the production), and also to network effects such as homophily and network correlation (for the consumption).

Finally, we examine the variance of the production and consumption polarities. We ask whether users who are more partisan also present a *lower* variance in their polarities, which means they produce and consume content from a narrower spectrum of sources. Figure 4 shows the consumption and production variance of each user (y-axis) against the respective (mean) polarity measure. The plot shows a clear "downward U" trend, which confirms the aforementioned hypothesis: bipartisan users follow news sources with a wider spread of political leaning, rather than just picking from the center, which makes their news diet qualitatively different from partisan users. We obtain similar results when looking at the variance of production and consumption polarities as a function of user polarity score [6] (omitted due to space constraints). The consistency of these results reinforces the validity of our production and consumption polarity metrics.

5.2 Analysis of partisan users

Recall that a δ-partisan user is one who produces content exclusively from one side of the political spectrum. In this section, we study how partisan users differ from bipartisan users. We focus on three main elements for the comparison:

(a) *Network*: PageRank (global measure of importance), clustering coefficient (local measure of community connection), and absolute user polarity (higher values indicate higher polarization).

(b) *Profile*: number of followers (proxy for popularity), number of friends, number of tweets (proxy for activity), age on Twitter (number of weeks the user has been on Twitter).

(c) *Interaction*: retweet/favorite rate, retweet/favorite volume.

Partisans and bipartisans are parameterized by a threshold δ, and we consider different values for δ between 0.20 and 0.45 in steps of 0.05. For each value of δ, we explore the value distribution of the above features for the two groups of users and test whether they are different. Table 2 (second column) summarizes the results for partisan users and lists the features for which the difference is statistically significant on a majority of the datasets. A "✓" in the table means that the property (e.g., PageRank) is significantly higher for partisans for at least 4 of the 6 values of the δ threshold, for most of the datasets (In most cases we find consistent behavior across all datasets).[9] A "✓ (-)" means that the property is significantly lower for partisans. A "✗" indicates we find no statistically significant difference.

Table 2: Comparison of various features for partisans & bi-partisans and gatekeepers & non-gatekeepers. A ✓ indicates that the corresponding feature is significantly higher for the group of the column ($p < 0.001$) for at least 4 of the 6 thresholds δ used, for most datasets. A minus next to the check-mark (-) indicates that the feature is significantly lower.

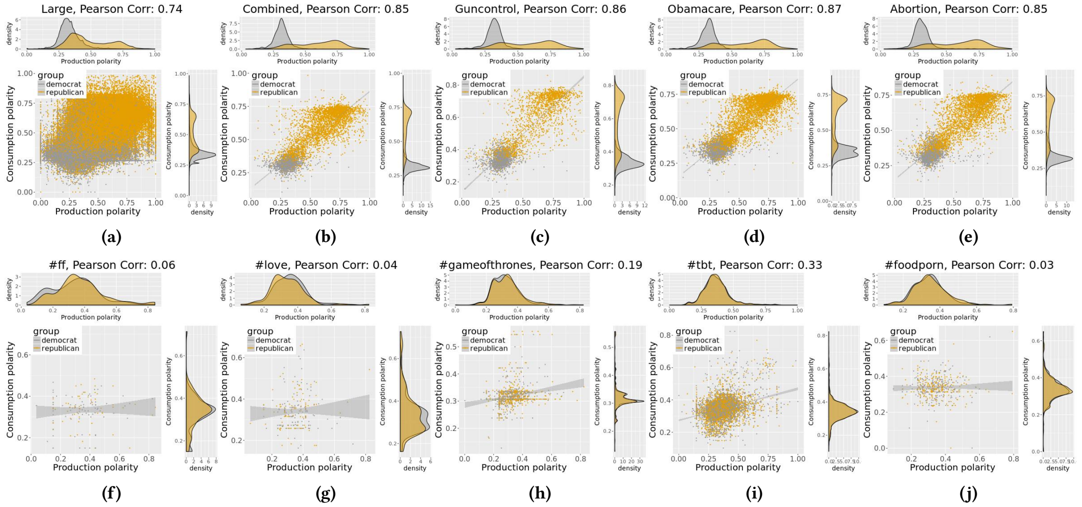
Features	Partisans	Gatekeepers
PageRank	✓	✓
clustering coefficient	✓	✓ (-)
user polarity	✓	✓ (-)
degree	✓	✓
retweet rate	✓	✗
retweet volume	✓	✗
favorite rate	✓	✗
favorite volume	✓	✗
# followers	✗	✗
# friends	✗	✗
# tweets	✗	✗
age on Twitter	✗	✗

For some of the features that exhibit significantly different distributions between the two groups, the distributions are shown in Fig. 5 (user polarity), Fig. 6 (PageRank), and Fig. 7 (clustering coefficient). Each figure shows a set of beanplots,[10] one for each Political dataset. Each beanplot shows the estimated probability density function for a measure computed on the dataset, the individual observations are shown as small white lines in a one-dimensional scatter plot, and the mean as a longer black line. The beanplot is divided into two groups, one for partisan users (left/dark) and one for bi-partisan ones (right/light).

Considering absolute polarity polarities, partisan users are significantly more polarized than bipartisan ones, as shown in Figure 5. We see that partisan users enjoy a more central position in the network, indicated by higher PageRank (Figure 6). Similarly, partisan users are more connected to their own community, indicated by a higher clustering coefficient (Figure 7). Finally, their tweets are more appreciated, i.e., a higher fraction of their tweets receives a retweet, albeit the effect size is smaller in this case (figure not shown). Similar trends hold for the number of retweets and the number of favorites (omitted due to space constraints). These results are consistent irrespective of the value of the δ threshold used to define δ-partisan users. We do not find any consistent trend across datasets in terms of profile features (Table 2).

[9] Significance tested using Welch's t-test for equality of means ($p < 0.001$) [36].

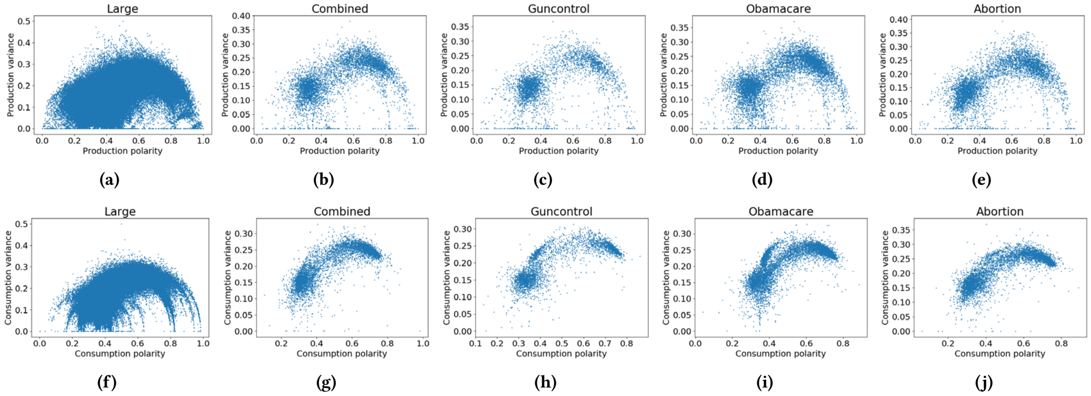
[10] A beanplot is an alternative to the boxplot for visual comparison of univariate data among groups.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 3: Distribution of production and consumption polarity, for Political (first row) and Non-Political (second row) datasets. The scatter plots display the production (x-axis) and consumption (y-axis) polarities of each user in a dataset. Colors indicate user polarity sign, following [6] (grey = democrat, yellow = republican). The one-dimensional plots along the axes show the distributions of the production and consumption polarities for democrats and republicans.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 4: Top: Production polarity variance vs. production polarity (mean). Bottom: Consumption polarity variance vs. consumption polarity (mean).

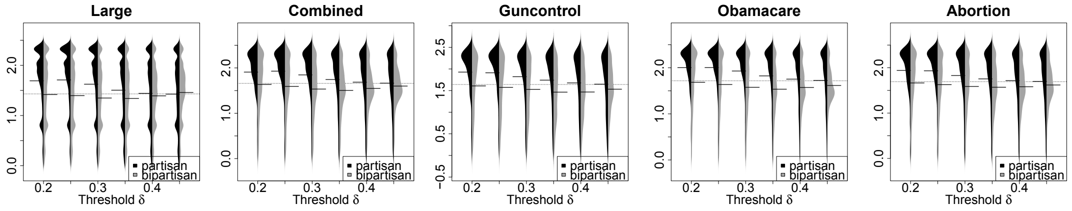
5.3 Gatekeepers of information

We now turn our attention to δ-gatekeeper users, i.e., users who consume more central content than what they produce. As in the previous section, we vary δ between 0.2 and 0.45 in intervals of 0.05 and compare gatekeepers with other users who are not gatekeepers. Due to space constraints, we do not show beanplots for the gatekeepers. We only show a summary of results in Table 3.

Gatekeepers, like partisans, occupy positions with high centrality in the network, i.e., higher than average PageRank and in-degree. However, differently from the rest of the side they align with, they show a lower clustering coefficient, an indication that they are not completely embedded in a single community. Given that they receive content also from the opposing side, this result is to be expected: most of the links that span the two communities will remain open (i.e., not form a triangle). Similarly, their polarity score is on average less extreme than the rest of their group.

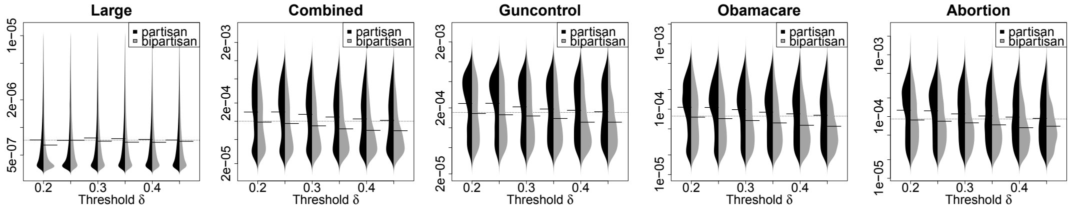
Differently from the partisans, we could not find consistent trends for interaction features such as retweet and favorite rate and volume. Profile features are also not consistently different for gatekeepers. The results are reported in Table 2.

Finally, given that both partisans and gatekeepers sport higher centrality, we compare their PageRank values directly and find that



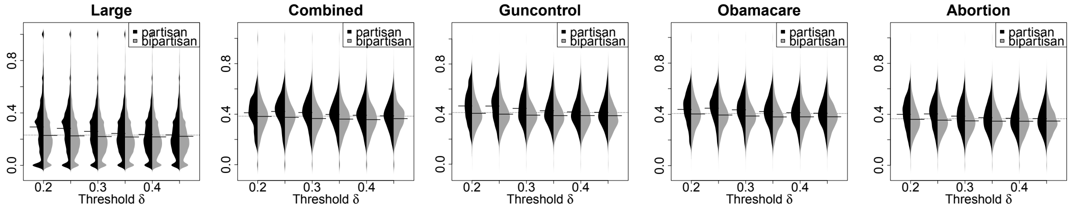
(a) (b) (c) (d) (e)

Figure 5: Absolute value of the user polarity scores for δ-partisan and δ-bipartisan users.



(a) (b) (c) (d) (e)

Figure 6: Pagerank for δ-partisan and δ-bipartisan users.



(a) (b) (c) (d) (e)

Figure 7: Clustering Coefficient for δ-partisan and δ-bipartisan users.

Table 3: Comparison between δ-gatekeeper users and a random sample of normal users. A ✓ indicates that the corresponding property is significantly higher for gatekeepers ($p < 0.001$) for at least 4 of the 6 thresholds δ used. A minus next to the checkmark (-) indicates that the property is significantly lower.

	PageRank	Degree	CC	Polarity
guncontrol	✓	✓	✓ (-)	✓ (-)
obamacare	✓		✓ (-)	✓ (-)
combined	✓	✓	✓ (-)	✓ (-)
abortion	✓	✓	✓ (-)	✓ (-)
large	✓	✓	✓ (-)	✓ (-)

there is a significant difference: partisans have a higher PageRank compared to gatekeepers (figure not shown). This effect is more pronounced for higher values of the threshold δ, possibly suggesting that, even among users who produce polarized content, purity (not following users of the opposite side) is rewarded.

5.4 Prediction

Given that partisans and gatekeepers present markedly different characteristics in terms of network and content, can we predict who is who without looking at their production and consumption polarities? That is, how evident is their role in the discussion just by looking at their network, and profile features? We train a Random Forest classifier on the Political datasets, and use the following features for each user:

- *Network features*: PageRank, degree, clustering coefficient;
- *Profile features*: number of tweets, of followers, of friends, age on Twitter;
- *Tweet features*: *n*-grams with tf-idf weights from their tweets.

We fix an intermediate threshold $\delta = 0.3$ to define the set of partisans and gatekeepers for each dataset. We build balanced classification tasks by picking an equal number of partisans/gatekeepers and a random sample of non-partisan/non-gatekeeper users.

The accuracy of the classification model is shown in Table 4 (average for 10-fold cross-validation) for partisans (p) and gatekeepers

Table 4: Accuracy for prediction of users who are partisans (*p*) or gatekeepers (*g*). (net) indicates network and profile features only, (*n*-gram) indicates just n-gram features. The last two columns show results for all features combined.

	p (net)	*g* (net)	*p* (*n*-gram)	*g* (*n*-gram)	*p*	*g*
combined	0.71	0.67	0.73	0.65	0.81	0.67
guncontrol	0.70	0.64	0.76	0.62	0.83	0.67
obamacare	0.75	0.65	0.78	0.64	0.83	0.66
abortion	0.71	0.63	0.76	0.65	0.80	0.69
large	0.72	0.70	0.74	0.68	0.78	0.75

(*g*). Given that the classification datasets are balanced, a random guess would have an accuracy of 0.5. However, all features give a better prediction. It is interesting to see that just using simple n-gram features performs well. This hints that there are marked differences in the way partisans and gatekeepers use text. Note that n-gram features, even though using content, are not related to the production/consumption polarity computation, as these scores are only computed using tweets with links to news sources (and not the actual content itself). Identifying partisans shows to be markedly easier than gatekeepers, with accuracies hovering around 80% for partisans compared to 70% for gatekeepers, when using all features combined. Therefore, we conclude that being a partisan has clear correlations with specific network and content features that enable their identification with high accuracy.

6 DISCUSSION

In this paper we study echo chambers in political discussions in social media, in particular, we study the interplay between content and network, and the different roles of users. Germane to our approach is the definition of measures for the political leaning of content shared by users in social media. These measures, which are grounded on previous research [4], capture both the leaning of the content shared by a single user, as well as the leaning of the content to which such user is exposed, by virtue of its neighborhood in the social network.

Characterising echo chambers. When applied to discussions about politically contentious topics, our results support the existence of political echo chambers. In particular, the distribution of production and consumption polarities of users is clearly bi-modal, and the production and consumption polarities are highly correlated. Conversely, the phenomenon does not manifest itself when the topic of discussion is not contentious. This result reinforces the validity of the proposed measures — and agrees with similar conclusions presented by Barberá [6], where retweet networks exhibit higher polarization for political topics.

Partisan users. We highlight the "price of bipartisanship" in terms of various aspects, including network position, community connections, and content endorsement. Overall, bipartisan users pay a price in terms of network centrality, community connection, and endorsements from other users (retweets, favorites). This is the first study to show the price of being bipartisan, especially in the context of political discussions forming echo chambers. This result highlights a worrying aspect of echo chambers, as it suggests the existence of latent phenomena that effectively stifle mediation between the two sides.

Gatekeepers. Finally, we examined gatekeepers, i.e., users who are bipartisan consumers but partisan producers. These users lie in-between the two opposed communities in network terms, but side with one in content terms. Their clustering coefficient is usually lower, as they have links to both communities, which are unlikely to be closed. The role of gatekeepers has not been examined in the context of echo chambers. Previous studies on Twitter showed that gatekeepers are typically ordinary citizens [37] rather than officially active partisans (e.g., party members).

We also experimented with a different definition of gatekeepers – users who have a high consumption variance and low production variance. This definition captures a slightly broader set of users (compared to Equation 5), e.g., users who consume from both ends of the political spectrum and produce balanced 'centrist' content. The results were almost similar to the ones reported above in Section 5.3, and so we do not present them explicitly.

Nevertheless, from our current analysis, it is not clear if such users act as open-minded net-citizens or "sentinels" who want to be informed about and attack the opinions of the opposition. Given the importance such users appear to have in the network structure (higher PageRank, and higher indegree (more followers)), this aspect remains to be studied in future work. In the former case (i.e., if gatekeepers are open-minded net-citizens), gatekeepers would be good candidates for users to nudge towards the opposing side [16, 28]. The possibility of identifying gatekeepers to a non-random extent by just using network features (e.g., if they do not actively produce content) makes an interesting application.

Limitations. As with any empirical work, this study has its limitations. First, the datasets used are just a sample of all the discussions in social media, and they all come from Twitter. However, 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. We have tried to address concerns the generality of our results by performing our analysis on datasets of various sizes, from various domains and time periods.

Second, our production and consumption scores rely on external labeling of news sources along a political axis. This choice limits the applicability of our analysis to debates that are politically aligned, and mostly for English-speaking and US-related topics. This limitation is not inherent in the methodology, but is due simply to the availability of such data. Media bias and labeling of media on a political axis is a field in itself (media and communication studies), and hence, this is not a big limitation. See [20] for a review on media bias and different ways to label media sources.

Moreover, our analysis assumes that each user consumes all content produced by all their neighbors. That is, we use the "follow" relationship as a proxy for the actual content consumption. In reality, a user might not consume everything from people they follow. In the absence of scroll or click logs, which could give us more fine-grained results, this proxy is the best we can get.

Finally, it is possible that not all news articles from the news sources we base the polarity measures are political. During preprocessing, we attempted to split news articles from these sources into hard (politics, opinion, etc.) and soft news (gossip, entertainment, etc.) and applied the classifier from Bakshy et al. [4]. We found that almost all (over 85%) of the urls from these domains are classified as hard news — and so, we opted to consider all of them in our analysis, knowing that a small fraction of them might not be "hard" political news.

Future work. The results shown in this study are just one step towards the understanding of echo chambers, which open up several directions for future work.

First, exploring more nuanced content and network features, which might lead to a better understanding of echo chambers in social media. For instance, *n*-gram features turned out to be very informative for identifying partisans, which indicates a distinctive writing style of this set of users. In this study we focused on content polarity based on a ground truth, but more powerful NLP techniques (e.g., topic modeling) might enable more powerful analysis.

Second, designing (probabilistic generative) models to capture the observed echo-chamber structure in terms of content and network features. Our findings 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 [5]. However, in light of the current 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.

Acknowledgments. This work was supported by the Academy of Finland projects "Nestor" (286211), "Agra" (313927), and "AIDA" (317085), and the EC H2020 RIA project "SoBigData" (654024).

REFERENCES

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

[2] Jisun An, Daniele Quercia, Meeyoung Cha, Krishna Gummadi, and Jon Crowcroft. 2014. Sharing political news: the balancing act of intimacy and socialization in selective exposure. *EPJ Data Science* 3, 1 (2014), 12.

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

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

[5] Sven Banisch and Eckehard Olbrich. 2017. Opinion Polarization by Learning from Social Feedback. *arXiv preprint arXiv:1704.02890* (2017).

[6] Pablo Barberá. 2015. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis* 23, 1 (2015), 76–91.

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

[8] Karine Barzilai-Nahon. 2009. Gatekeeping: A critical review. *Annual Review of Information Science and Technology* 43, 1 (2009), 1–79.

[9] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and information technology* 15, 3 (2013), 209–227.

[10] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. In *ICWSM*.

[11] Daniel DellaPosta, Yongren Shi, and Michael Macy. 2015. Why do liberals drink lattes? *Amer. J. Sociology* 120, 5 (2015), 1473–1511.

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

[13] Arthur Edwards. 2013. (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 (2013), 127–150.

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

[15] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying Controversy in Social Media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)*. ACM, New York, NY, USA, 33–42. DOI: https://doi.org/10.1145/2835776.2835792

[16] Kiran Garimella, Aristides Gionis, Nikos Parotsidis, and Nikolaj Tatti. 2017. Balancing Information Exposure in Social Networks. In *Neural Information Processing Systems (NIPS)*.

[17] Kiran Garimella and Ingmar Weber. 2017. A Long-Term Analysis of Polarization on Twitter. In *Proceedings of the 11th AAAI International Conference on Web and Social Media*. AAAI.

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

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

[20] Tim Groeling. 2013. Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science* 16 (2013).

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

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

[23] K Hazel Kwon, Onook Oh, Manish Agrawal, and H Raghav Rao. 2012. Audience gatekeeping in the Twitter service: An investigation of tweets about the 2009 Gaza conflict. *AIS Transactions on Human-Computer Interaction* 4, 4 (2012), 212–229.

[24] Eric Lawrence, John Sides, and Henry Farrell. 2010. Self-segregation or deliberation? Blog readership, participation, and polarization in American politics. *Perspectives on Politics* 8, 1 (2010), 141–157.

[25] Kurt Lewin. 1943. Forces behind food habits and methods of change. *Bulletin of the national Research Council* 108, 1043 (1943), 35–65.

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

[27] Haokai Lu, James Caverlee, and Wei Niu. 2015. Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In *CIKM*. ACM, 213–222.

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

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

[30] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

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

[32] Xiaoyan Qiu, Diego FM Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. 2017. Limited individual attention and online virality of low-quality information. *Nature Human Behavior* 1 (2017).

[33] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. 2016. Echo Chambers on Facebook. (2016).

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

[35] 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*.

[36] Bernard L Welch. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika* 34, 1/2 (1947), 28–35.

[37] Weiai Wayne Xu and Miao Feng. 2014. Talking to the broadcasters on Twitter: Networked gatekeeping in Twitter conversations with journalists. *Journal of Broadcasting & Electronic Media* 58, 3 (2014), 420–437.

Aalto-DD 20/2018

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.

9 789526 078328

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

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS