CONTROVERSY ANALYSIS AND DETECTION

A Dissertation Presented

by

SHIRI DORI-HACOHEN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2017

Computer Science

CONTROVERSY ANALYSIS AND DETECTION

A Dissertation Presented

by

SHIRI DORI-HACOHEN

Approved as to style and content by:

_____

James Allan, Chair

_____

Justin Gross, Member

_____

David Jensen, Member

_____

W. Bruce Croft, Member

_____

Laura Dietz, Member

_____

James Allan, Chair of the Faculty
Computer Science

DEDICATION

To Gonen and Gavahn

ACKNOWLEDGMENTS

First and foremost, I'd like to thank my advisor, James. Thank you for your immense support throughout my Ph.D. and its many twists and turns. From the day I tentatively knocked on your door asking for a project, you were always honest and kind with your feedback, holding me to a high standard while always implying that I was definitely up to the task.

Funding wise, this work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #IIS-0910884, in part by NSF grant #IIS-1217281 and in part by Google. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect those of the sponsor.

I'd like to thank multiple friends for numerous thoughtful conversations and commenting on various drafts of my papers throughout the years. Elif, Marc, Jeff and Henry provided oft-needed guidance in my early years; along with CJ, David W., Jiepu, Laura, Mostafa, Nada, Weize, Zeki and all my labmates at CIIR: I was fortunate to count you not only among my colleagues but also as friends. The CIIR and CICS staff helped make my research go smoothly: thanks Barb, Dan, Kate, Jean, Joyce, Leeanne, and Vickie! To my co-authors: David J., Elad, James, John, and last but not at all least Myung-ha - it was a pleasure working with you.

Multiple people, too numerous to mention by name, have been helpful in my journey to study the computational aspects of controversy, whether by patiently discussing the topic with me, pointing me to relevant articles or news stories, or sharing valuable resources. My thanks go to all of you! In particular, I am extremely grateful to Allen Lavoie, Hoda Sepehri Rad, Taha Yasseri and Elad Yom-Tov for shaping my thinking through fruitful discussions

and sharing resources early during my Ph.D. Thanks also go to Brendan O'Connor for sharing his rich Twitter data set, and to Wikimedia and Pew Research for making their data sets publicly available and easily accessible.

Thanks to CICS CS Women, CRA-W and Grace Hopper Celebration for providing a much-needed community in my early years; to Kathryn S. McKinley, Angela Demke Brown and Kim Hazelwood for encouraging conversations and the existence proof; to Alexandra, Brian and Emery of CICS for their support; to the GWIS co-founding members; to Jackie and Bobby for leading the way; to Laura Dietz, my WIR co-conspirator; and to Diane Kelly for a remark that changed my career path.

Special thanks to Raz Alon and Myung-ha Jang for their invaluable support and assistance during the final stages of this dissertation, and to Janice Plado Dalager for her incredible help throughout the last year of my Ph.D. To my SuperBetter allies and my RBT friends, too many to be individually named, and especially to my two accountability partners, Chamika & Dave: I owe you an immense debt of gratitude for your endless encouragement. To my UMass cohort: Cibele, Fabricio, Ravali, Sandeep: thanks for walking the path with me! To my faithful friends from elsewhere: Ala, Ariel, Aruna, Celal, Einat, Elior, Ilana, Itay, Marcel, Moitrayee, Moran, Nava, Noa, Omer, Ornit, Tidhar, Tsahi, Yael, Yousef: thanks for always being there for me even from half the world away! Limi: you kept helping when the going got tough. Thanks also to Nina for staunch support over 5 years, to Anna for help at crucial junctures, and to Rob for helping me cross the finish line.

Thanks to Birton Cowden, Steve Willis, Karen Utgoff and Matt Wallaert for helping me discover my entrepreneurial path, and to David Jensen, Justin Gross and Brian Levine for reflecting back my excitement about my work's potential impact.

David Allen, Brené Brown, Carol Dweck, Josh Kaufman, Margaret Lobenstine and Cal Newport, among many others, wrote books that changed my life and allowed me to unleash my potential on the poor, unsuspecting world. Ramit Sethi challenged me to do so, and also created a community in which I found my home. Shalini Bahl, Kim Nicol and

Corinne Andrews were the best spiritual teachers one could hope for. My sincere gratitude extends to all of them.

Gavahni, thank you for being patient with me while I was off “getting my doctor costume”. Your presence, smiles, hugs and budding sense of humor made the last three and a half years a huge joy. I’m so blessed to have you in my life!

Gonen: none of this would have happened if not for your post-doc and our conversation in D.C. during the summer of 2010. Words fail to express my appreciation and admiration for everything you’ve sacrificed to help me get here - not to mention the need to listen to my endless complaints. Well, we always did say that one should do what one is good at. Thanks for helping me figure out what that was and then support me while doing it. You and Gavahn make all the effort worthwhile.

ABSTRACT

CONTROVERSY ANALYSIS AND DETECTION

SEPTEMBER 2017

SHIRI DORI-HACOHEN

B.Sc., UNIVERSITY OF HAIFA

M.Sc., UNIVERSITY OF HAIFA

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor James Allan

Seeking information on a controversial topic is often a complex task. Alerting users about controversial search results can encourage critical literacy, promote healthy civic discourse and counteract the "filter bubble" effect, and therefore would be a useful feature in a search engine or browser extension. Additionally, presenting information to the user about the different stances or sides of the debate can help her navigate the landscape of search results beyond a simple "list of 10 links". This thesis makes strides in the emerging niche of controversy detection and analysis. The body of work in this thesis revolves around two themes: computational models of controversy, and controversies occurring in neighborhoods of topics. Our broad contributions are: (1) Presenting a theoretical framework for modeling controversy as contention among populations; (2) Constructing the first automated approach to detecting controversy on the web, using a KNN classifier that maps

from the web to similar Wikipedia articles; and (3) Proposing a novel controversy detection in Wikipedia by employing a stacked model using a combination of link structure and similarity. We conclude this work by discussing the challenging technical, societal and ethical implications of this emerging research area and proposing avenues for future work.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

The Internet, the Web, and technologies for information retrieval have massively expanded information access for billions of people over the past twenty years. Social network tools such as Twitter, Facebook, discussion forums, and comments on news articles [Dori-Hacohen and Shavit, 2013] are increasingly the place where democratic arguments are being held. Technological tools hold an increasingly crucial role in shaping these discussions by influencing which users see which data, through algorithmic curation and filtering. Publishing material about controversial issues is of paramount importance to a functioning democratic society, because it allows our disagreements to be aired in public. However, when searching for discussion of a controversial issue it is all too easy to cherry-pick from the results. For example, those against gun rights will surely find material supporting this position (the tragedy of school shootings), whereas those for gun rights will find other evidence (the Second Amendment in the U.S.). Meanwhile, alternative medicine sites appear alongside pediatrician advice websites, the phrase "global warming is a hoax" is in wide circulation, and political debates rage in many nations over economic issues, same-sex marriage and healthcare.

Unfortunately, critical literacy, civic discourse and trustworthy information are not immediate results of effective information retrieval. Access does not always translate into trustworthy information: e.g., parents seeking information about vaccines will find plenty of "proof" that they cause autism, and may not even realize the depth of the controversy involved [Walia, 2013]; ads for helplines displayed to users searching for "abortion" are discreetly funded by pro-life (anti-abortion) religious groups [Heroic Media, 2014]. The

underlying thread connecting all these examples is that users searching for these topics may not even be aware that a controversy exists; indeed, without the aid of a search engine feature or browser extension to alert them, they may never find out. Even if they are aware of the controversy's existence, the challenge of navigating the different sides of the debate, and understanding which website lies on what side, remains a cognitive burden on the user.

Several researchers have claimed that search engines have significant political power. Introna and Nissenbaum explicitly called out the politics of search engines as shaping the web [Introna and Nissenbaum, 2000]. In his book *Republic.com 2.0*, legal scholar Cass Sunstein argues that a purely consumer-based approach to Internet search is a major risk for democracy [Sunstein, 2009]. One of deliberative democracy's basic tenets, he argues, is the ability to have a shared set of experiences, and to be exposed to arguments you disagree with. Controversies proliferate online; yet search engines and social media are increasingly responsible for "Filter Bubbles", wherein click-feedback and personalization lead users to only see what they want, serving to further increase confirmation bias [Pariser, 2011, Sunstein, 2009]. While this may seem to match individual users' preference, the net effect on society is potentially detrimental. There is preliminary evidence showing that exposure to diverse opinions can improve civic discourse [Yom-Tov et al., 2013]. Given the existence of filter bubbles, the benefits of exposure to diverse opinions will only be available to users who can detect controversial topics. We discuss the need for controversy detection in greater detail in Section 1.1.

Unfortunately, the current state of affairs is that we simply do not understand controversy well enough from a computational perspective. Algorithms based on incomplete understanding are bound to fail in a variety of unexpected ways, replicating or even exacerbating the sources of human bias in the data. Recent work on controversy cuts across traditional disciplinary lines to include a wide variety of computational tasks along with social science and humanities [Dori-Hacohen et al., 2015], and has made significant strides in analyzing and detecting controversy (cf. [Garimella et al., 2016, Borra et al., 2015]).

Nonetheless, serious gaps remain in our theoretical and practical understanding of how to define controversy, and how it manifests and evolves. For example, polling organizations naturally segment their results based on population groups such as race and gender, but these notions are surprisingly absent from algorithmic analyses of online data. Instead, controversy is assumed to be an absolute, single value for an amorphous global population.

Meanwhile, a disparity is growing between scientific understanding and public opinion on certain controversial topics, such as climate change, evolution, or vaccines [Leshner, 2015], with many scientists explicitly fighting these trends by insisting "there is no controversy", referring to *scientific* controversy (cf. [Helfand, 2016]). Still, non-scientific claims and arguments continue to proliferate, raising exposure to the (supposedly non-existent) controversies. As researchers studying controversies online, how are we to reconcile the oft-repeated argument from the scientific community that "there is no controversy" with the practical appearance of wildly diverse opinions on said topics? In other words, is climate change controversial[1]?

Additionally, it is becoming increasingly evident (including through our work) that algorithmically recognizing controversy is a challenging task, and that users searching for controversial topics should be presented with additional information beyond a standard Search Engine Results Page (SERP) with matching documents [Dori-Hacohen et al., 2015]. We are interested in techniques that encourage and facilitate healthy debates, allowing users to critically approach these issues. We believe that informing users about controversial topics would be a valuable addition to the end-user experience; this requires detecting such topics as a prerequisite. For example, imagine an alert presented at the top of a web page: "This webpage represents one of several perspectives on a controversial topic." To do so, we need to answer a non-trivial question: "Is this topic controversial?" Automated tools

[1]This differs from a value judgment, such as "*Should* climate change be controversial?".

performing such detection can support users in their browsing and search experience [Dori-Hacohen and Allan, 2015, Dori-Hacohen et al., 2015].

1.1 The Need for Controversy Detection and Analysis

In two provocative books, journalist Eli Pariser and legal scholar Cass Sunstein argue that increasing personalization reduces exposure to diverse opinions, creating a serious risk for deliberative democracy and the tenets it relies on [Pariser, 2011, Sunstein, 2009]. Search engines and social media use personalization to tailor results to the users' opinions, a phenomenon termed the "Filter Bubble" [Pariser, 2011], which can further exacerbate confirmation bias. In one study on user evaluation of web pages, only 11.6% of users noted the bias present in the information as part of their evaluation of a website's credibility, trailing far after attributes such as design look (mentioned by 46.1%) or information structure (28.5%) [Fogg et al., 2003]. Information has a clear effect on the choices people make, such as shifting their voting patterns [DellaVigna and Kaplan, 2007] or affecting their medical outcomes [Yom-Tov and Boyd, 2014]. Disputed information, such as Barack Obama's birthplace supposedly being Kenya, proliferates on the web [Dong et al., 2015]. Wikipedia is a valuable resource, but often "hides" the existence of debate by presenting even controversial topics in deliberately neutral tones [Wikipedia, 2014], which may be misleading to people unfamiliar with the debate. For example, the Wikipedia article for U.S. President Barack Obama make him appear non-controversial, while both the Talk page and an automated analysis show otherwise.

As a result of these concerns, a computational approach to controversy detection and analysis is an area of growing interest in the intersection of several areas of computer science, such as information retrieval, social media analysis, computational social science, natural language processing, argument mining and trustworthiness. Computational analysis of controversy in the web, news and social media holds exciting and important implications for civil discourse and critical literacy [Yom-Tov et al., 2013], yet is replete with

technical challenges. For example, accurately and automatically distinguishing between controversial and noncontroversial topics is one such relevant challenge which is currently within technical reach, yet is far from a solved problem.

This thesis focuses on solving a variety of open problems within the nascent field of computational controversy detection and analysis, and advancing the state of the field. Prior to embarking on this thesis, there was little work in the space of controversy detection and analysis, and a lack of clarity as to the goals and challenges arising from this new and exciting area. This gap stems from the distance between the social need for better approaches to handle controversy, and the little prior work that existed on controversy from a computational perspective. We will discuss this prior work in Chapter 2. We will also reflect further on challenges and potential implications of computational analysis of controversies when we discuss opportunities for work beyond this thesis, in Chapter 7.

1.2 Technical Overview

This thesis positions controversy detection and analysis as a useful and achievable goal of search engines. We further the state of the art in the field with regards to detection and analysis of controversial topics. Due to our research, it is now possible for a search engine to inform its users with some level of reliability that their query or webpage discusses a controversial topic even if the page itself might appear staid and reliable. Addressing the gap present in the literature at the outset of this thesis, we make both technical and conceptual contributions to the emerging niche, or subfield, of controversy detection and analysis. The body of work in this thesis revolves around two themes: computational models of controversy, and controversies occurring in neighborhoods of topics.

The first theme introduces computational models by which to understand controversy. In Chapter 3, we define "contention", a measure of disagreement that is rooted in the notion of populations in addition to topics. This work draws on insights from the social sciences in order to present a theoretical framework that hypothesizes "contention" as one

crucial dimension of controversy. We validate our model by examining a diverse set of sources: real-world polling data sets, actual voter data, and Twitter coverage on several topics. We also present evaluation on 2000 Wikipedia topics for contention. Additionally, we demonstrate how one previous work on controversy in Wikipedia, the heuristically-based **M** score [Sumi et al., 2011], can be understood as an approximation of "contention". We demonstrate that the contention measure holds explanatory power for a wide variety of observed phenomena, such as controversies over climate change and other topics that are well within scientific consensus. Finally, we re-examine the notion of controversy, and present a theoretical framework that defines it in terms of population. We present preliminary evidence showing that contention is only one dimension of controversy, along with "importance" and perhaps others such as "conviction". Our new contention measure, along with the hypothesized model of controversy, suggest several avenues for future work.

As part of the evaluation of contention described above, we demonstrate that the Wikipedia **M** score can be recast as an instantiation of our contention model (see section 6.1). Next, in Chapter 4, we use **M** as well as other Wikipedia scores in a new task: a binary classification of whether a web page discusses a controversial topic. We leverage the Wikipedia scores to identify controversial topics on the web using a k-Nearest-Neighbor classifier, thus tying into our second theme of topical neighborhoods of controversy.

Much of the limited prior work on controversy detection has focused on Wikipedia (see Chapter 2), and mostly on analyzing each page in isolation or with regards to its editors. This thesis examines several results emerging from an alternative hypothesis, which is the second theme of this thesis: that controversies occur in neighborhoods of related topics. Our work on classifying web documents (Chapter 4) relied on this hypothesis implicitly. In Chapter 5, we examine controversy in Wikipedia articles directly and find that they exhibit homophily with regards to controversy. In other words, related topics are more likely to have similar levels of controversy. We then demonstrate that this homophily can

be leveraged to improve controversy detection on Wikipedia pages, using a novel algorithm based on techniques of collective inference and stacked models.

Finally, in Chapter 6, we tie these strands of research together by introducing an instantiation of our probabilistic contention score on Wikipedia. We compare this score to a heuristic score from prior work, and evaluate its ability to classify controversy in Wikipedia both quantitatively and qualitatively. We then briefly evaluate the ability to use this contention score for the extrinsic task of detecting controversy on the web.

1.3 Contributions

In this thesis, we make the following technical contributions:

1.3.1 Contributions regarding the definition of controversy (Chapter 3):

1.3.1.1 Re-conceptualizing controversy

We re-conceptualize the term "controversy" and offer a new theoretical framework to understand it. Our framework departs from most existing work about controversy in two major ways. First, in contrast to prior work which considers controversy to have a single global value, we define controversy not only in terms of its topic, but also in terms of the population being observed. This yields different controversy scores for different populations regarding the same topic. Second, in contrast to prior work which considers controversy as single-dimensional, we define controversy as multi-dimensional. We present preliminary evidence suggesting that the dimensions of controversy include contention (defined in the remainder of the chapter), alongside other dimensions, such as "importance".

1.3.1.2 Defining contention

We define a novel quantitative measure we call "contention", which captures disagreement and is likewise defined with respect to a topic and a population. As in contribution 1.3.1.1 above, this measure departs from prior work by yielding different contention scores

for a given topic depending on the population observed. In addition, our measure also depart from prior work by accounting for participants in the population who hold no stance with regards to a specific topic, as well as allowing for any number of stances rather than just two opinions. We model contention from a mathematical standpoint and validate our model by examining a diverse set of sources: real-world polling data sets, actual voter data, and Twitter coverage on several topics.

1.3.1.3 Explanatory power of our framework

We demonstrate that the re-conceptualized controversy framework and the contention measure hold explanatory power for a wide variety of observed phenomena that cannot be explained under previous global controversy views:

1. Controversies over climate change, vaccines, and other topics that are well within scientific consensus, and which scientists often say "there is no controversy". These can be explained under the new model: there is indeed no controversy within the scientific community, while there is still controversy among the general population in certain regions.

2. International conflict (such as the Israeli-Palestinian conflict) can be understood as exhibiting high contention at the global level, often with moderate to low contention within each participating nation.

3. Well-documented polling variations in controversy among certain populations or interest groups, such as different attitudes toward corporal punishment among different racial groups, can be easily modeled under population-dependent contention.

4. Topics that are controversial only in certain geographical regions or among certain interest groups can likewise be modeled.

1.3.1.4 Releasing Twitter data set

We release a Twitter data set of nearly 100 million tweets, for several popular topics in the last eighteen months, including three prominent controversies (the 2016 U.S. Elections, the UK referendum on leaving the EU, commonly known as Brexit, and "The Dress", a photo that went viral when people disagreed on its colors).

1.3.2 Contributions for classifying controversy on the web (Chapter 4):

1.3.2.1 First algorithm for detecting controversy on the web

We pose the novel problem of web classification of controversy (detecting controversial topics on the web) [Dori-Hacohen and Allan, 2013], and construct the first algorithm addressing it. Our algorithm is based on a K-Nearest-Neighbor classifier that maps from webpages to related Wikipedia articles, thus leveraging the rich metadata available in Wikipedia to the rest of the web.

1.3.2.2 Human-in-the-loop approach to controversy classification

We demonstrate that using a human oracle for determining controversy in Wikipedia articles can achieve an $F_{0.5}$ score of 0.65 for classifying controversy in webpages. We show absolute gains of 22% in $F_{0.5}$ on our test set over a sentiment-based approach, highlighting that detecting controversy is more complex than simply detecting opinions.

1.3.2.3 Fully automated web classification of controversy

We construct a fully automated system for web classification of controversy that relies on automated scoring of Wikipedia articles. We demonstrate that our system is statistically indistinguishable from the human-in-the-loop approach it is modeled on, and achieves similar gains over prior work baselines (20% absolute gains in $F_{0.5}$ measure and 10% absolute gains in accuracy).

1.3.2.4 Releasing web and Wikipedia data set

We collect and release a data set of 377 web pages and 1761 Wikipedia articles annotated with regards to controversy, which is the first data set available for this new problem, and the largest data set of controversy labels to date. The data set also includes 3430 annotations of pairs of webpages and Wikipedia articles, regarding whether or not the Wikipedia page is on the same topic as the webpage.

1.3.3 Contributions for collective classification of controversy on Wikipedia (Chapter 5):

1.3.3.1 Wikipedia articles exhibit homophily with respect to controversy

We demonstrate that Wikipedia articles exhibit homophily with respect to controversy. In other words, pages that are linked on Wikipedia are more likely than random to have the same controversy label ($p < 0.001$).

1.3.3.2 Collection inference algorithm for controversy detection in Wikipedia

We present a novel algorithm for controversy detection in Wikipedia, based on techniques of collective inference and stacked models, that leverages the homophily demonstrated above. This approach used a combination of link structure and similarity to find "neighbors" and rank them, and is comparable to the state of the art for this problem. We evaluated our approach on the data set released above, as well as on another data set available from prior work [Sepehri Rad et al., 2012].

1.3.3.3 Sub-network approach using similarity

We present a new sub-network approach, that uses similarity to select neighbors for the stacked model. This approach is not limited to the controversy problem domain, and can be used in other problem areas in which homophily is present with semi-structured data sets.

1.3.3.4 Neighbors-only classifier

We present a neighbors-only classifier that does not utilize the features of a page itself but only on its neighbors, and demonstrate that, counter-intuitively, in certain cases it can be as effective as a classifier that relies on the page's own features.

1.3.4 Contributions for contention on Wikipedia and the web (Chapter 6):

1.3.4.1 Revert-based contention for Wikipedia

We define a variation of the contention model which can be applied to Wikipedia, based on special types of edits called reverts, and demonstrate that the M score from prior work [Yasseri et al., 2014] is an approximation of contention.

1.3.4.2 Evaluating contention and M on Wikipedia

We evaluate our probabilistic score on a set of 2000 Wikipedia articles, and find that it yields similar results to the M score. We also briefly evaluate the effects of using the contention score on the web data set from Chapter 4. Finally, we perform a qualitative evaluation on the differences between the scores by examining the top 1000 articles using each score.

Finally, we end the thesis with a discussion of implications of our work, including the technical, social and ethical challenges that arise from computational controversy detection, and conclude with several avenues for future work.

We now turn to describe related work and how it is distinct from our work, and then proceed to discuss and describe these contributions in detail in the following chapters.

CHAPTER 2

RELATED WORK

Several strands of related research inform our work: the definition of controversy, the need for controversy detection, controversy detection in Wikipedia, controversy on the web and in search, fact disputes and trustworthiness, as well as arguments, stances and sentiment. In some of our work, we also use approaches for collective classification. We describe each area in turn, and highlight our own prior work in the area, referring to sections of the thesis discussing these contributions.

2.1 What is the definition of Controversy?

How does one define controversy? While there is no one definition of the term controversy, we might use the following definition as an approximation: controversial topics are those that generate strong disagreement among large groups of people. Like the definition of relevance, it's possible that controversy should be defined operationally: whatever people perceive as controversial, is controversial.

However, in line with others' findings [Klenner et al., 2014], our research so far shows that achieving inter-annotator agreement on the "controversy" label is very challenging. Additionally, while intuition and some researchers might suggest that the notion of sentiment should be relevant for controversy (e.g. [Popescu and Pennacchiotti, 2010, Tsytsarau et al., 2011]), others have argued that sentiment is not the right metric by which to measure controversy [Awadallah et al., 2012b, Dori-Hacohen and Allan, 2013, Mejova et al., 2014]; opinions on movies and products may contain sentiment, yet lack controversy. These concerns have led us to search for a formal definition of controversy.

Research on controversies in computer science has nearly universally considered controversy as either a binary state or a single quantity, both of which are to be measured or estimated directly [Awadallah et al., 2012a, Sepehri Rad and Barbosa, 2012, Borra et al., 2015]. Prior work almost exclusively did not model controversy formally, but rather classified controversy based on various choices of ground truth or else based on implicit definitions of controversy. There are two recent exceptions that modeled controversy directly, one of them by this author: Amendola et al. [2015] used 5-star movie rankings and modeled two types of controversy, which they call "hard" and "soft" controversy, based on the distribution of star rankings [Amendola et al., 2015]. In our recent work[1], we offered the first formal model of controversy in textual documents [Jang et al., 2016]. Even in these rare instances where controversy was modeled formally, the meaning of the term "controversy" was not put forth as a question, but assumed to be a known quantity in the world. Most prior work in computer science does not define controversy at all, and treats it as a global quantity (cf. [Kittur et al., 2007, Yasseri et al., 2014]).

Likewise, we find some of the definitions of controversy used by others, or the data sets that those definitions lead them to use, to be very problematic. In one early paper in the field, the definition for controversy conflated vandalism and controversy, and therefore rated "podcast" as the most controversial topic in Wikipedia [Vuong et al., 2008], along with other pages such as that for celebrity actress "Emma Watson" which is a highly vandalized page, but not controversial. Another paper relies on the list of Lamest Edit Wars in Wikipedia as a controversy data set [Bykau et al., 2015], a list which includes topics such as whether Caesar Salad was named for Julius Caesar or for restauranteur Caesar Cardini. This aforementioned list article is topped by the warning: "This page contains material which is kept because it is considered humorous. Please do not take it seriously."[2]We suspect the choice of this list as ground truth to train a classifier for controversy demonstrates

[1]Which is not part of this thesis.

either a lack of cultural understanding or of the satirical nature of the said list. (We further discuss disagreement on frivolous topics, and its connection to controversy, below.) In Chapter 3, we depart from prior work in computer science by focusing on a more achievable goal of measuring what we call "contention", a population-dependent measure, and offering a mathematical framework to define it while grounding it in empirical data.

Meanwhile, most of the work on controversy in social studies and humanities is qualitative by nature, and often focuses on one or two examples of controversy (c.f. [Szívós, 2005, Van Eemeren and Garssen, 2008]), or else works towards a more qualitative analysis of the overall patterns across controversies [Dascal, 1995]. One notable exception [Cramer, 2011] used word occurrence and frequency of controversy-related noun-phrases in several corpora such as the Reuters corpus [Rose et al.,] to construct an analysis of the use of the term "controversy". For example, corpus-wide usage of determiners ("the controversy") and adjectives or qualifiers before or after the term "controversy" were tabulated in order to differentiate between different types of controversies. Cramer also searched for occurrence of specific terms as **controversy**, **dispute**, **scandal**, and **saga** and qualitatively studied their context. Cramer explains that "controversy" cannot necessarily be verified to exist in the world independent of its appearance in text, but rather it is created and shaped by the discourse surrounding it, particularly in news outlets [Cramer, 2011].

In philosophy, Leibniz offered a simple definition of controversy: a controversy is a question over which contrary opinions are held [Leibniz, 1982], a definition which Dascal notes as "clearly insufficient" [Dascal, 1995]. Dascal offers a theory of controversies which distinguishes between types of polemic discourse [Dascal, 1995]. Dascal's distinctions between "controversy", "discussion" and "dispute" depend on whether people share an underlying worldview and on whether they are trying to convince each other vs. a third party. Cramer explicitly refrains from defining the term "controversy" directly, referring

[2]https://en.wikipedia.org/wiki/Wikipedia_talk:Lamest_edit_wars

to it as a "metadiscursive" and "indexical" term, and says it can be loosely defined as *something that you would know when you see it* [Cramer, 2011].

Though one may have an intuitive understanding of the term "controversy", without a structured definition, our work (as well as others') will not hold as much weight or predictive power. In Chapter 3, we depart from prior work by forgoing the notion of a single, universal controversy score for a topic. Rather, we introduce a novel measure we call "contention", defined with respect to both a topic and *a population*. We model contention from a mathematical standpoint and validate our model by examining a diverse set of sources. We demonstrate that the contention measure holds explanatory power for a wide variety of observed phenomena, such as controversies over climate change and other topics that are well within scientific consensus. Our model for contention draws on insights from existing computational, humanities and social sciences work, yet departs from it by offering a formal computational model for "contention", and offers a re-conceptualization of controversy.

Chen and Berger, while discussing whether controversy increases buzz and whether that is good for business, propose that "controversial issues tend to involve opposing viewpoints that are strongly held" [Chen and Berger, 2013]. However, these definitions leave a gap when people disagree on opinions that are strongly held on frivolous topics such as the colors of a dress[3], the proper orientation of toilet paper, or on the various topics included in the Lamest Edit Wars list mentioned above [Bykau et al., 2015]. Likewise, one may inquire whether the scope and context of the controversy matters, and whether the user performing the search is relevant. For example, do controversies regarding occurrences on American Idol (which may induce edit wars on Wikipedia) matter less than a controversy on the Israeli-Palestinian Conflict? One could argue that the latter is a much more controversial and influential topic; but for the user searching for "American Idol" or, for example, "Joanna Pacitti" (a controversial contestant on the show), perhaps the knowledge that this represents a controversial topic may be just as relevant – in the context of that

search. Humans intuitively understand that disagreements over toilet paper or the color of "The Dress" are qualitatively different from "serious" or "important" controversies such as the Israeli-Palestinian conflict, Brexit or the supposed connection between vaccines and autism. Yet this gap remains completely unexplained in the existing literature. In our work, we present the first explanation for this gap when we hypothesize controversy as a population-dependent, multidimensional quantity, for which "contention" and "importance" are possible dimensions (see Section 3.2.2), and present preliminary evidence for this hypothesis. To the best of our knowledge, this model is the first to account for two real-world phenomena: (1) The "importance" dimension is the first attempt to account for the gap between disagreement on frivolous topics vs. serious ones, and (2) The population-dependent aspect of our model is the first attempt to explain how a topic can be controversial in the population group for which it is salient, while being irrelevant to the rest of the world.

2.2 Controversy Detection in Wikipedia

Of the relatively sparse prior work on automatically detecting controversy, most focuses on automatically detecting article-level controversial topics in Wikipedia, a task originally proposed by Kittur et al. [Kittur et al., 2007]. Wikipedia's collaborative nature, along with its versioning system, is a rich resource and a natural focus for an investigation of controversy, since its rich user-generated content base offers a wealth of semi-structured data that can be mined (cf. [Kittur et al., 2007, Sepehri Rad and Barbosa, 2012, Sumi et al., 2011, Yasseri et al., 2012]). Kittur et al. proposed a logistic regression classifier based on several metrics such as the length of the article and its associated talk page, the number of users, and so forth [Kittur et al., 2007]. Several studies of the topic focused on who edits whom (e.g. [Brandes and Lerner, 2008, Jesus et al., 2009]), though some researchers

[3] A topic we analyze in detail in Chapter 3.

mistakenly conflated vandalism with controversy [Vuong et al., 2008]. Sepehri Rad et al. presented a binary classifier for controversy using collaboration networks [Sepehri Rad et al., 2012], and also presented a comparative study of the various approaches to detecting controversy in Wikipedia [Sepehri Rad and Barbosa, 2012]. Another approach used the article feedback tool 5-star ratings as a signal for controversy to detect controversy at the article level [Jankowski-Lorek et al., 2014]. Sumi, Yasseri and colleagues proposed a hand-crafted, heuristic score of controversy, which they call M, that is based on the notion of Edit Wars and mutual reverts [Sumi et al., 2011, Yasseri et al., 2012]. Reverts are a special kind of edit in Wikipedia, whereby a user completely undoes some previous edit; the proposed score was based on the number of editors who reverted each other, and their presumed reputation. However, their validation of the score was limited, and a later comparative study argued that this score did not hold enough discriminative power as a classifier [Sepehri Rad and Barbosa, 2012]. We will examine the M score further and give it a more theoretical underpinning in Section 6.1.

One consistent feature among the wide diversity of approaches in the prior work regarding Wikipedia, is that nearly all the papers mentioned above use an approach that classifies each page in isolation [Bykau et al., 2015, Jankowski-Lorek et al., 2014, Kittur et al., 2007, Yasseri et al., 2012]. In contrast, our work examines networks of Wikipedia articles that are topically related, and argues that controversy occurs in neighborhoods - a major theme in this thesis (see Chapters 4 & 5). While some recent work has alluded to the possibility that controversies occur in neighborhoods of related topics [Das et al., 2013] or demonstrated such clusters anecdotally [Jesus et al., 2009], this potential connection had yet to be tested or used to improve controversy detection. We first explored this theme implicitly in our work on detecting controversy on the web (Chapter 4), which leveraged similarity between Wikipedia and web pages to detect controversy. We then turn to explicitly testing whether controversy indeed runs in neighborhoods in Chapter 5, where we shall

demonstrate that Wikipedia articles exhibit homophily with respect to controversy (Section 5.1), and then utilize this attribute in order to improve classification accuracy (Section 5.2).

Aside from a page-level classification of controversy, other related tasks have also been examined in the context of Wikipedia. In a seminal paper from Wikipedia's early days, the collaboration and conflict that are revealed through Wikipedia's edit history were visualized powerfully [Viégas et al., 2004]. Another study demonstrated that dispute discussions on the Wikipedia Talk pages can be classified successfully using a sentence-level sentiment analysis approach [Wang and Cardie, 2014]. Fine grained analyses of Wikipedia edits at the word or sentence level have also been examined [Borra et al., 2015, Bykau et al., 2015], and sentence-level visualization of controversy "hot spots" proposed [Borra et al., 2015].

Detecting controversy in Wikipedia is an important challenge, and can be seen as an end in itself, as described above. That said, these detection methods have a wider reach, and can be used as a step for solving other problems. Das et al. explored the possibility of Wikipedia administrators attempting to manipulate controversial pages in a certain area by extending any controversy metric into a clustered controversy measure, hypothesizing that editors focusing on a certain controversial topic were more invested in the outcomes than those spreading their edits across several topical areas [Das et al., 2013]. Wikipedia has been used in the past as a valuable resource assisting in controversy detection elsewhere, whether as a lexicon [Popescu and Pennacchiotti, 2010] or as a hierarchy for controversial words and topics [Awadallah et al., 2012b]. In our work on detecting controversy in the web (see Chapter 4), we use automatic query generation as a bridge between the rich metadata available in Wikipedia and the sparse metadata on the web; we thus demonstrate that controversy detection in Wikipedia is an effective proxy for detecting controversy on the web, which can in turn can be used as labels in a search engine or browser extension serving end users.

2.3 Controversy on the Web and in Search

One of our goals in Chapter 4 is to widen the scope of controversy detection past Wikipedia to the entire web. Outside of Wikipedia, none of the scaffolding or rich meta-data used by this research exists for webpages, and therefore these algorithms cannot be directly applied to detect controversy elsewhere.

Some targeted domains such as news [Awadallah et al., 2012b, Choi et al., 2010] and Twitter [Popescu and Pennacchiotti, 2010] have been mined for controversial topics, mostly focusing on politics and politicians. Choi et al. created a mixture model of topics and sentiment, extracting controversial topics (noun or verb phrases) at a sentence level; their focus was mostly on generating an appropriate list of sub-topics for a controversial topic. Awadallah et al. built a corpus of opinions expressed by politicians on controversial topics, and leveraged Wikipedia in a different manner in order to create a network of controversial topics and subtopics [Awadallah et al., 2011, Awadallah et al., 2012b]. Theirs and other work relied on domain-specified sources such as Debatepedia[4] [Awadallah et al., 2012b, Kacimi and Gamper, 2012] that are politics-heavy. Debate websites often focus exclusively on political issues; as of this writing Debatepedia has no entry discussing Homeopathy. Therefore, these approaches do not generalize to non-political controversies; we depart from these papers by approaching all controversies, whether political, medical, or religious, and in all web pages.

Popescu & Pennacchiotti constructed three supervised models for detecting controversial events in Twitter, focusing in particular on "celebrity" entities [Popescu and Pennacchiotti, 2010]. However, they treat "controversy" as a simple binary variable: an event is either controversial or not, and the meaning of the term "controversy" is implicit (as discussed above). In Chapter 3, we use Twitter to validate our "contention" measure; however, we depart from this work in several ways, by focusing on a formal definition of contention.

[4]Debatepedia: http://dbp.idebate.org/

In some cases, a simple word search can be useful in detecting controversial queries [Gyllstrom and Moens, 2011]; unfortunately, this query-side approach to controversy detection relied on the Google Suggest API, which was deprecated in 2011 as part of the Google Labs shutdown. Assuming one knows that a query is controversial, diversifying search results based on opinions is a useful feature [Kacimi and Gamper, 2012]. We consider the problem of detecting controversy to have potential utility as a precursor step in diversifying controversial queries, though that is not our main focus in this work.

When reading about claims that may be in dispute, users do not actively seek contrasting information or viewpoints, but they are more likely to read the contrasting viewpoint when it was explicitly portrayed as such [Vydiswaran et al., 2012]. Yom-Tov et al. described how, for polarized political topics, two versions of seemingly similar search queries existed; using one query or the other exposed the biases of the user issuing the query (e.g. "obamacare" vs. "affordable health care") [Yom-Tov et al., 2013]. Their research demonstrated that interspersing search results from the opposite viewpoint on such polarized queries had the double effect of both encouraging users to read more diverse opinions, and to read more news in general [Yom-Tov et al., 2013]. Prior work has shown that there is value in presenting users with information that may differ from their original perspective, whether it is portrayed implicitly or explicitly [Ennals et al., 2010, Vydiswaran et al., 2012, Yom-Tov et al., 2013]. This partially counteracts the "filter bubble" effect [Pariser, 2011]. This is the motivation behind our work in Chapter 4: to automatically inform users of the various viewpoints on controversial topics.

In our work, we introduce the problem of detecting controversy in the web, and propose two solutions to it (see Chapter 4). To the best of our knowledge, this problem has not been formulated as such before, though several special cases have been explored by previous researchers (e.g. on Wikipedia or other targeted domains, as discussed above). We use a k-nearest-neighbor approach in order to leverage the existing work on detecting controversy

in Wikipedia, connecting Wikipedia articles to the web via a query-generation approach, as described in Chapter 4.

2.4 Fact disputes and trustworthiness

Fact disputes and the analysis of trustworthiness are often related to controversial topics [Ennals et al., 2010, Vydiswaran et al., 2012]. Similar to our goal, the Dispute Finder tool focused on finding and exposing disputed claims on the web to users as they browse [Ennals et al., 2010]. However, Dispute Finder was focused on manually added or bootstrapped fact disputes, not on controversies at large. Some controversies indeed originate from arguments over fact disputes (such as the birthplace of Barack Obama) or over matters of scientific study (the effect of vaccines on autism). In other cases, however, judgments, values and interpretation can fuel the fire despite widespread agreement on the basic facts of the matter; moral debates over abortion and euthanasia, or political debates over the size of the U.S. government and the Israeli-Palestinian conflict, are some examples. Recent work has also suggested rating websites based on the trustworthiness based on the facts they include [Dong et al., 2015]. Recent concerns over fake news have also raised this topic to public awareness. In contrast to this body of work, we are interested in scalably detecting controversies that may stem from fact disputes, but also from disagreement on values or from moral debates. We discuss the connection between controversies and fact disputes further in Chapter 7.

2.5 Arguments, stances and sentiment

Sentiment analysis can naturally be seen as a useful tool as a step towards detecting varying opinions, and potentially controversy [Choi et al., 2010,Popescu and Pennacchiotti, 2010, Tsytsarau et al., 2011, Cartright et al., 2009]. Sentiment-based diversification of search on controversial topics has been proposed as well [Aktolga and Allan, 2013, Kacimi and Gamper, 2012]. Some might argue that sentiment analysis should be used for this

problem domain (see, e.g., [Wang and Cardie, 2014]) on the grounds that controversy is likely to engender strong sentiment, which is one reason we compare to sentiment baselines [Aktolga and Allan, 2013] in some of our work (Chapter 4).

However, sentiment analysis and opinion mining work has long focused on product or movie reviews, where users have little intention or reason to hide their true thoughts; in contrast, many biased webpages intentionally obfuscate the sentiment and controversy involved, attempting to pass their vision as objective. Therefore, we hypothesize that detecting controversial webpages requires a dedicated approach that will likely not be based solely on sentiment analysis, and must rely on information outside the text of the page itself - at the very least, other websites on the same topic. Controversy is also subtler than sentiment, as described by Awadallah et al.: "controversies are much more complex and opinions are often expressed in subtle forms, which makes determining pro/con polarities much more difficult than [existing] work on opinion mining" [Awadallah et al., 2012b].

Other work, including our own (see Chapter 4), has likewise shown that sentiment and controversy are overlapping, but far from identical, constructs; and additionally demonstrated that sentiment analysis is not sufficient to detect controversy [Dori-Hacohen and Allan, 2013, Mejova et al., 2014], though it may be useful as a feature [Popescu and Pennacchiotti, 2010]. Likewise, polarity only gives partial information about how controversial topics are [Klenner et al., 2014]. Sentiment analysis is likely to be more effective when considering its variance in analyzing online conversations, such as the dispute discussions on the Wikipedia Talk pages [Wang and Cardie, 2014], Twitter [Garimella et al., 2016], or forum discussions [Hasan and Ng, 2013], rather than when examining individual Wikipedia articles or webpages.

Assuming one has successfully discovered that a document or topic is controversial, another challenge (beyond the scope of this thesis) is understanding what is controversial about it. In the political sphere, Awadallah and colleagues demonstrated automatic extraction of politician opinions [Awadallah et al., 2012b]; however, their work relied heavily

on news sources. The diversificiation research mentioned above could be seen as a form of stance extraction [Kacimi and Gamper, 2012]. While frameworks for machine-readable argumentation and "The Argument Web" have been implemented (see, e.g. [Bex et al., 2014]), search engines cannot rely on widespread adoption of such tools. Borra et al. [Borra et al., 2015] demonstrated an algorithm that detects which topics are most contested within a given Wikipedia page. Recently, a stance detection algorithm based on random walks was proposed for Twitter [Garimella et al., 2016]. The research in this area has interesting applications to our work, particularly regarding automatically detecting stance groups or presenting users with explicit stances on controversial topics, which are beyond the scope of this thesis but extremely interesting related questions. In Chapter 3, we use a high-precision, low-recall manually curated approach to create stance groups. We also discuss future work in stance detection in Chapter 7.

2.6 Web-page Classification and General Collective Classification Approaches

In Chapter 5, we use a modified version of collective inference; we survey key related work here. Collective and relational inference are machine learning techniques that can be applied to relational data. Collective inference can be applied in relational data sets where there are dependencies known or presumed between related instances, and particularly on the special case of web-page classification. Collective inference has been demonstrated to be successful on many complex problems such as hyperlink categorization [Chakrabarti et al., 1998] and Wikification [Cheng and Roth, 2013], by exploiting autocorrelation between objects that are related [Fast and Jensen, 2008, Jensen et al., 2004].

Stacked models are a type of collective classification that avoids the need for computationally intensive inference procedures, and are particularly useful in situations where there is a lack of extensive ground truth data for the neighborhood of a page [Kou and Cohen, 2007]. In stacked models, an *intrinsic classifier*—one that relies only on the features of the

data instance being evaluated—is trained first. Subsequently, the intrinsic model is applied to generate predictions for the neighbors of every document in the set; these predictions are then aggregated and used as features of that document, in an extended data set. Finally, a *stacked model* is trained by using this extended data set, as in regular collective inference. In other words, the collective inference classifier is "stacked" over the intrinsic classifier. The chief difference is that instead of using known truth labels of neighbors, a stacked model uses the outputs of an intrinsic classifier. Stacked models have been demonstrated to be effective at collective classification due to a reduction in bias [Fast and Jensen, 2008].

When stacked models are used in semistructured data sets (given their historical roots in the knowledge discovery community), they are usually applied in a relational manner [Abiteboul et al., 2000, pg v]: the notion of relatedness is often defined directly by relations available in the structured data set. However, as demonstrated in several domains, a simple relational link between two objects does not necessarily imply a strong connection between them (see, e.g. citation strength in scholarly articles [Dietz et al., 2007]). Though similarity has been used in a few instances as an enhancing feature to improve collective inference and label propagation (cf. [Bröcheler et al., 2012] [Wang and Sukthankar, 2013]), most papers use relational links alone. In one case, Kuwadekar and Neville developed an approach called *relational active learning*, which takes similarity into account when choosing new examples to label [Kuwadekar and Neville, 2011].

Inspired by these papers and by the needs of our Wikipedia classification task (see Section 5.2), we explicitly construct a network with a subset of the edges based on similarity, which is not explicit in the relational data, for the purpose of improving stacked classification. Rather than considering the relations in the data set as fixed, we use features of a semi-structured data set, such as directionality of relations and similarity between objects, to construct a more useful notion of relationship. In this, we depart from most stacked classification approaches that tend to assume that the data set contains a relational schema which is fixed (cf. [Fast and Jensen, 2008, Jensen et al., 2004, Kou and Cohen, 2007]).

Our work in Chapter 5 is also distinct from Probabilistic Similarity Logic [Bröcheler et al., 2012], since they propose a first-order logic approach to reasoning about similarity for inference purposes, whereas we propose to construct an induced subgraph of relationships based directly on similarity measures.

2.7 Summary

Since the computational study of controversies is fairly new, this thesis draws on a wide variety of prior work from the social sciences and humanities, as well as the small yet rapidly growing number of studies in computer science examining various aspects of controversies. The related work and the gaps in it point to a deep need to define controversy formally, which we tackle in Chapter 3. We also use our formal definition to re-derive a particular heuristic from prior work, the M score [Yasseri et al., 2012], and compare our probabilistic, theoretically-motivated derivation to the original score (Section 6.1). We draw on prior work in Wikipedia to formulate a new task of detecting controversy on the web (see Chapter 4), based on the implicit expectation that controversy runs in topical neighborhoods. We then explicitly examine this hypothesis in a Wikipedia data set in Chapter 5, and propose a modification to state-of-the-art stacked classifiers in order to leverage this homophily.

CHAPTER 3

MODELING CONTROVERSY AS CONTENTION WITHIN POPULATIONS

3.1 Introduction

As discussed in Chapter 2, a growing body of research focuses on computationally detecting controversial topics and understanding the stances people hold on them, yet gaps remain in our theoretical and practical understanding of how to define controversy, how it manifests, and how to measure it.

In this chapter, we address these gaps by proposing a theoretical framework for controversy, drawing on insights from social science and humanities and marrying them with the mathematical rigor of a computational approach. Our theoretical framework holds two major departures from the existing work about controversy. First, we define controversy not only in terms of its topic, but also in terms of the population being observed. Second, we conceive of controversy as composed of at least two dimensions, rather than being a one-dimensional quantity. We then proceed to examine one of these dimensions, which we refer to as "contention", and model it rigorously from a mathematical standpoint. In an additional departure from most past work, our contention model allows for any number of stances rather than just two.

These elements give our model explanatory power that can be used to understand a large variety of observed phenomena, ranging from international conflict, through community-specific controversies, as well as the aforementioned high-stakes public opinion controversies over scientifically well-understood phenomena such as climate change, evolution, and vaccines.

In order to ground our theoretical model, we examine a diverse collection of data sets from both online and offline sources. First, we examine several real-world polling data sets, among them a poll that focuses on opinions about scientific topics, such as climate change and evolution, measured among the general U.S. population as well as the scientific community [Pew Research Center, 2015a, Pew Research Center, 2015b]. Additionally, we look at Twitter coverage for several popular topics in the last eighteen months, including three prominent controversies (the 2016 U.S. Elections, the UK referendum on leaving the EU, commonly known as Brexit, and "The Dress", a photo that went viral when people disagreed on its colors). We cross-reference contention from Twitter with other data sources: a popular online poll for "The Dress", and actual voter data for Brexit and the U.S. Elections.

Our new contention measure, along with the hypothesized model of controversy, afford new directions of understanding of controversy, such as the growth of contention or controversy over time among different populations, and points to open questions for future research.

3.2 Reconceptualizing Controversy

We mathematically formulate a model of controversy based on a notion of a population and the people within it. We suggest an approach which rather than modeling controversy directly, focuses on modeling amounts of disagreement or "contention". A certain level of contention may or may not meet criteria for controversy, depending on other features of the controversy model.

3.2.1 Preliminary definitions

Let $\Omega = \{p_1..p_n\}$ be a population of n people. Let T be a topic of interest to at least one person in Ω.

We propose to re-conceptualize controversy in a way that is inseparable from the population we are observing.

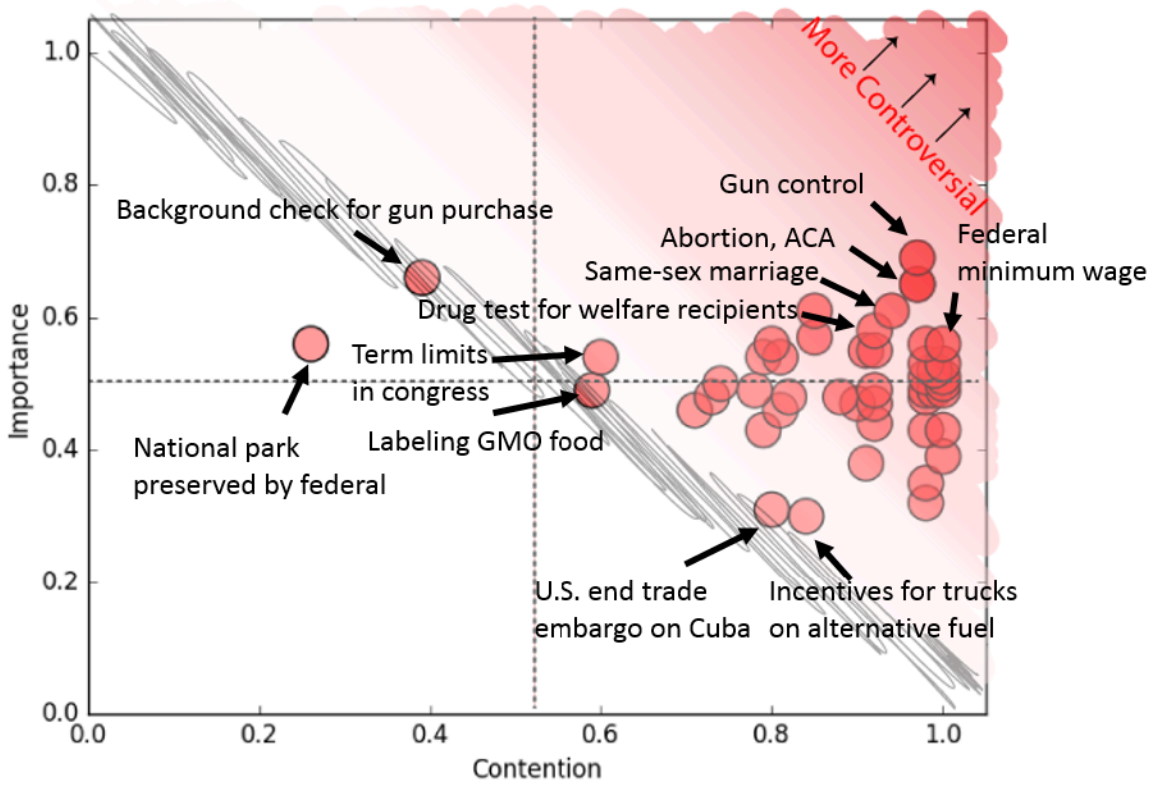

Figure 3.1: iSideWith topics plotted reconceptualizing controversy as composed of at least two dimensions, contention and importance. Sample topics are given in each quadrant of {low,high} importance and contention.

We thus define the level of controversy with respect to a topic and a group of people: Let $controversy(\Omega, T)$ represents the level of controversy of topic T within Ω.

3.2.2 Controversy is Multidimensional

Consider the cases of the Brexit referendum and "The Dress", two controversies which we will explore in further detail below. When observed among the population which considered them as salient, both were extremely contentious, in the sense that nearly any group of people sampled from these populations was strongly divided in their opinion. However, it is immediately obvious that placing Brexit and "The Dress" in the same bucket is somewhat problematic. One, a political referendum on Britain's decision whether to exit the European Union, affects the fate of entire nations, with far-reaching and difficult to pre-

dict implications on diplomatic relationships and the world economy for years to come. The other, an under-developed photo of a mother-of-the-bride's dress, caused a surprising divided reaction in color perception, went viral around the world, and was subsequently forgotten by nearly everyone. Its impact on the world was likely negligible, with the exception of a burst of scientific papers in visual perception studying this unexpected effect [Schlaffke et al., 2015, Journal of Vision Special Collection, 2016][1].

Therefore, we propose a new model in which controversy is composed of at least two orthogonal dimensions, which together play a role in determining how controversial a topic is for a given population, one of which is "contention". We can hypothesize other possible dimensions. For example, a possible second dimension is "conviction", i.e. encoding how strongly people hold their opinions [Chen and Berger, 2013]. However, this dimension is insufficient to explain such arguably frivolous controversies as "The Dress", toilet paper orientation [Wikipedia, 2016], or the Lamest Edit Wars in Wikipedia [Bykau et al., 2015]. An additional orthogonal metric is needed in order to distinguish between contention and controversy. Therefore, we hypothesize the existence of a notion of "importance" or "impact" as another possible dimension of controversy, which we believe to be minimally required in order to make sense of observed phenomena such as debates over "The Dress". Using the same notation as above, we hypothesize that these are minimal dimensions of controversy, though there may be others:

$$\begin{aligned} controversy(\Omega, T) = f(&contention(\Omega, T), \\ &conviction(\Omega, T), \\ &importance(\Omega, T)...) \end{aligned}$$

This framework is demonstrated schematically with two dimensions in Figure 3.1, overlaying actual results including "importance" as reported in the iSideWith data set (see Ta-

[1] And now, this thesis.

ble 3.1). The first dimension is "contention" which we will define shortly, and measures the proportion of people who are in disagreement. The other dimension is "importance", which we loosely define as the level of impact of that issue to the world, and which was self-reported by users of iSideWith. In Figure 3.1, we hypothesize controversy to be a two-dimensional concept. An issue is more controversial when it has high contention and high importance (i.e., towards the right upper corner of Figure 3.1). Figure 3.1 shows a quadrant where an issue can have a {high, low} contention with a {high, low} importance. Issues such as gun control, abortion, and the affordable care act have high contention and high importance, hence are more controversial. Issues such as whether the government should provide incentives for trucks to run on alternative fuels is highly contentious, but is rated by users as low importance. Likewise, whether National parks should be preserved by the federal government is rated as somewhat important, but not contentious. We can consider issues that have only high contention or only high importance to be qualitatively different from each other, and both are overall less controversial than the issues that have both high importance and high contention. Using this framework, we can understand the disparity between "The Dress" and Brexit: the former is contentious with low importance (lower right quadrant), and thus not as controversial as Brexit with its high contention and high importance (upper right quadrant). Likewise, the Lamest Edit Wars list[2] includes a plethora of low importance topics that nonetheless generate contention on Wikipedia.

While computationally exploring the additional hypothesized dimensions, "conviction" and "importance" is left for future work, we have demonstrated that contention clearly is one such dimension, and that at least one additional dimension is required in order to fully understand controversies. We also hypothesize the existence of "importance" as an additional, orthogonal dimension of controversy. Contention does not fully capture the nuances of what we intuitively understand to be controversial, and adding the orthogonal

[2]https://en.wikipedia.org/wiki/Wikipedia_talk:Lamest_edit_wars

"importance" dimension adds further explanatory power to our framework. For the rest of this chapter, we focus our attention on modeling "contention" computationally.

3.3 Modeling Contention

We now proceed to formally model contention, the dimension of controversy which quantifies the proportion of people in disagreement within a population. We begin with a general formulation of contention, and then describe a special case in which stances are assumed mutually exclusive.

As before, $\Omega = \{p_1..p_n\}$ is a population of n people, and T is a topic of interest. Let c denote the level of contention, which we also define with respect to a topic and a group of people: $P(c|\Omega, T)$ represents the probability of contention of topic T within Ω. Let $P(nc|\Omega, T)$ or $P(\neg c|\Omega, T)$ similarly denote the probability of non-contention with respect to a topic and a group of people, such that: $P(c|\Omega, T) + P(nc|\Omega, T) = 1$.

Let s denote a stance with regard to the topic T, and let the relationship $holds(p, s, T)$ denote that person p holds stance s with regard to topic T. Let $\hat{S} = \{s_1, s_2, ..s_k\}$ be the set of k stances with regard to topic T in the population Ω. We allow people to hold no stance at all with regard to the topic (either because they are not aware of the topic, or they are aware of it but do not take a stance on it). We use s_0 to represent this lack of stance. In that case, let

$$holds(p, s_0, T) \iff \nexists s_i \in \hat{S} \text{ s.t. } holds(p, s_i, T),$$

Let $S = \{s_0\} \cup \hat{S}$ be the set of $k + 1$ stances with regard to topic T in the population Ω. Therefore, $\forall p \in \Omega$, $\exists s \in S$ s.t. $holds(p, s, T)$. Now, let *conflicts*: $S \times S \rightarrow \{0, 1\}$ be a binary function which represents when two stances are in conflict. Note that a person can hold multiple stances simultaneously, though no stance can be jointly held with s_0. We set conflicts$(s_i, s_i) = 0$.

Let **stance groups** in the population be groups of people that hold the same stance: for $i \in \{0..k\}$, let $G_i = \{p \in \Omega | holds(p, s_i, T)\}$. By construction, $\Omega = \bigcup_i G_i$. Let **opposing**

groups in the population be groups of people that hold a stance that conflicts with s_i. For $i \in \{0..k\}$, let $O_i = \{p \in \Omega | \exists j \text{ s.t. } holds(p, s_j, T) \wedge conflicts(s_i, s_j)\}$.

As a reminder, our goal is to quantify the proportion of people who disagree. Intuitively, we would like to have that quantity grow when the groups in disagreement are larger. In other words, if we randomly select two people, how likely are they to hold conflicting stances?

We model contention directly to reflect this question. Let $P(c|\Omega, T)$ be the probability that if we randomly select two people in Ω, they will conflict on topic T. This is equal to:

$$P(c|\Omega, T) = P(p_1, p_2 \text{ selected randomly from } \Omega, \exists s_i, s_j \in S, \text{ s.t. } holds(p_1, s_i, T) \wedge holds(p_2, s_j, T) \wedge conflicts(s_i, s_j))$$

Alternatively:

$$P(c|\Omega, T) = P(p_1, p_2 \text{ selected randomly from } \Omega, \exists s_i \in S, \text{ s.t. } p_1 \in G_i \wedge p_2 \in O_i).$$

Finally, we extend this definition to any sub-population of Ω. Let $\omega \subseteq \Omega, \omega \neq \emptyset$ be any non-empty sub-group of the population. Let $g_i = G_i \cap \omega$, and $o_i = O_i \cap \omega$. Thus, by construction, $g_i \subseteq G_i$ and $\omega = \bigcup_i g_i$. The same model applies respectively to the sub-population. In other words, for any $\omega \subseteq \Omega$,

$$P(c|\omega, T) = P(p_1, p_2 \text{ selected randomly from } \omega \wedge \exists i \text{ s.t. } p_1 \in g_i \wedge p_2 \in o_i).$$

3.3.1 Mutually exclusive stances

Note that we are selecting with replacement, and it is possible that $p_1 = p_2$. Strictly speaking, this model allows a person to hold two conflicting stances at once and thus be in

both G_i and O_i, as in the case of intrapersonal conflict. This definition, while exhaustive to all possible combinations of stances, is very hard to estimate. We now consider a special case of this model with two additional constraints. Let every person have only one stance on a topic:

$$\begin{aligned} \nexists p \in \Omega, s_i, s_j \in S \text{ s.t. } i \neq j \wedge \\ holds(p, s_i, T) \wedge holds(p, s_j, T). \end{aligned} \tag{3.1}$$

And, let every explicit stance conflict with every other explicit stance:

$$\textit{conflicts}(s_i, s_j) \iff (i \neq j \wedge i \neq 0 \wedge j \neq 0) \tag{3.2}$$

This implies that $G_i \cap G_j = \emptyset$. Crucially, we set a lack of stance not to be in conflict with any explicit stance. Thus, $O_i = \Omega \setminus G_i \setminus G_0$.

For simplicity, we estimate the probability of selecting p_1 and p_2 as selection with replacement[3]. Note that $|\Omega| = \sum_{i \in \{0..k\}} |G_i|$ and the probability of choosing any particular pair is $\frac{1}{|\Omega|^2}$. The denominator, $|\Omega|^2$, expands into the following expression:

$$|\Omega|^2 = (\sum_i |G_i|)^2 = \sum_{i \in \{0..k\}} |G_i|^2 + \sum_{i \in \{1..k\}} (2|G_0||G_i|) + \sum_{i \in \{2..k\}} \sum_{j \in \{1..i-1\}} (2|G_i||G_j|)$$

Depending on whether the pair of people selected hold conflicting stances or not, they contribute to the numerator in $P(c|\Omega, T)$ or $P(nc|\Omega, T)$, respectively. Therefore,

$$P(c|\Omega, T) = \frac{\sum_{i \in \{2..k\}} \sum_{j \in \{1..i-1\}} (2|G_i||G_j|)}{|\Omega|^2}$$

and

$$P(nc|\Omega, T) = 1 - P(c|\Omega, T) = \frac{\sum_{i \in \{0..k\}} |G_i|^2 + \sum_{i \in \{1..k\}} (2|G_0||G_i|)}{|\Omega|^2}$$

[3]The calculation is very similar for selection without replacement, except for extremely small population sizes.

As before, we can trivially extend this definition to any non-empty sub-population $\omega \subseteq \Omega$ using $g_i = G_i \cap \omega$. By construction, there is no contention within any single-stance group, g_i, with respect to topic T. In other words, $P(c|g_i, T) = 0$. Additionally, by construction, there is no contention within $g_i \cup g_0$, i.e. $P(c|g_i \cup g_0, T) = 0$.

By extension, if there is only one explicit stance s_1 with regard to topic T in the population Ω, there will be no contention in the population with respect to the topic. In other words, $|\hat{S}| \leq 1 \implies P(c|\Omega, T) = 0$.

3.3.2 Normalization factor

Trivially, $P(c|\omega, T)$ is maximal when when $|g_0| = 0$ and $|g_1| = ... = |g_k| = \frac{|\omega|}{k}$, and its value is $\frac{k-1}{k}$. This is subtly different from entropy due to the existence of s_0, as entropy would be maximal when $|g_0| = |g_1| = ... = |g_k| = \frac{|\omega|}{k-1}$.

Since the values of contention are in the range $[0, \frac{k-1}{k}]$ rather than $[0, 1]$, the probability scores will be sensitive to the number of stances k. To mitigate that effect, we can normalize the probability.

Since $\frac{k-1}{k}$ is the maximal contention score for k stances, let its inverse $n_k = \frac{k}{k-1}$ be the normalization factor. Now, let $nC = n_k * P(c|\omega, T)$ be the normalized contention score, which will now fall in the range $[0, 1]$. We can then define $nNC = 1 - nC$ as the normalized non-contention score. For example, a score of $P(c|\omega, T) = 0.2$ for 2 stances would result in normalized scores of $nC = \frac{2}{1} * 0.2 = 0.4$ and $nNC = 1 - 0.4 = 0.6$. The same score of $P(c|\omega, T) = 0.2$, if normalized for 3 stances, would instead yield normalized scores of $nC = \frac{3}{2} * 0.2 = 0.3$ and $nNC = 1 - 0.3 = 0.7$. This normalization brings both contention and non-contention to a full range of $[0, 1]$ each, with a contention score of $nC = 1$ signifying the highest possible contention, regardless of the total number of stances. For the remainder of this thesis, we use the normalized scores in place of the actual probability.

It's worth noting at this point that while our model theoretically allows for many stances, in practice the available evaluation data focuses almost exclusively on two stances. The two exceptions here are the AAAS data, in which one topic had three stances, and the U.S. election data, which we analyze in two variants: with or without third party candidates, i.e. 6 vs. 2 stances.

3.4 Data Collection and Preparation

In order to ground our model in empirical data, we collected several data sets. First, we collected data sets that represent explicit stance information, from informal online polls, through phone surveys, to actual voting records (on Brexit and the 2016 U.S. Elections). The complete set of explicit-stance data sets appears in Table 3.1. Between these data sets, we cover a wide variety of public opinion issues and a span of over 50 years. Second, we collected a set of tweets on several topics, one focusing on Brexit, and the other on "The Dress" phenomenon (see Table 3.2); in both, the stances taken by people are implicit and must be estimated.

3.4.1 Polling data sets

In the Pew and Gallup data sets, we used the topline survey results as reported by the respective organizations. For a given poll topic T, ω is the set of respondents, s_i are the set of response possibilities, and "no answer" represents s_0. This determines g_i and thus allows us to calculate $P(c|\omega, T)$ as above. In the case of statistically representative polls, conclusions can be generalized for the wider population from which the poll sample was drawn (within the margin of error of the polls).

Using one data set acquired from Pew Research Center, a non-partisan fact tank in the U.S., we are able to examine attitudes towards a number of issues among two populations: U.S. adults and U.S. scientists (Pew Adults and Pew AAAS in Table 3.1). The opinions for U.S. adults was gathered among a representative sample of 2,002 adults nationwide, while

Table 3.1: Data sets containing explicit stances. Survey types: A = Statistically Calibrated Phone Survey, B = Informal Online Polling, C = Public Voting Records.

Data set	Type	# Issues	Population(s)	Years	# People	Source
Gallup	A	3	US adults	1939-2016	varies (K)	[Gallup, 2016a, 2016b, 2016c]
Pew Adults	A	13	US adults	2014	2.0K	[Pew Research Center,
Pew AAAS	A	13	US scientists	2014	3.7K	2015a, 2015b]
iSideWith	B	52	US people	2014	varies (M)	By request
Buzzfeed	B	1	Online readers	2015-2016	3.5M	[Holderness, 2015]
Brexit Votes	C	1	UK voters	2016	46.5M	[The Electoral Comission, 2016]
U.S. Votes	C	1	U.S. voters	2016	251.1M	[McDonald, 2017; Wasserman, 2017; Wikipedia, 2017]

the opinions for scientists were gathered among a representative sample among the U.S. membership of the American Association for the Advancement of Science (AAAS) [Pew Research Center, 2015b].

We also obtained a data set from the iSideWith.com website, a nonpartisan Voting Advice Application [Cedroni, 2010] which offers users the chance to report their opinions on a wide variety of controversial topics, and outputs the information of which political candidate they most closely align with. We received the 2014 iSideWith data set by request from the website owners, which included nation-wide and per-state opinions over 52 topics. Each topic was posed as a question with two main options for answers, usually simply "yes" and "no". Additionally, the data set included the average importance of the issue (both nation-wide and per-state) rated by the users, which we use in our hypothesized controversy model (but not for contention).

3.4.2 Twitter data set

We collect a set of tweets on six events or topics from Twitter, which is available on our website[4]. We selected three contentious topics: "The Dress", the Brexit referendum, and the 2016 U.S. elections.

From the collected tweets, we identify two sub-groups of tweets by their stance revealed through their hashtags in order to measure their contention. In addition to the Twitter data, we also collected actual voting records for Brexit and the U.S. elections (see below for further description), as well as the Buzzfeed poll results for "The Dress" (see Table 3.1). For this purpose, we use the Twitter Garden Hose API, which allows us to collect 10% random sample of actual tweets if it is included in the sample.

"The Dress" refers to a photo that went viral over social media starting Feb. 26, 2015, after people couldn't agree on its colors. The photo was posted to tumblr and made popular by a Buzzfeed article asking "What color is this dress?" as a poll with two options, black and blue or gold and white; over 37 million people viewed the article to date [Holderness, 2015]. Over the course of the next 24 hours, "The Dress" made headline news in mainstream media outlets. The actual dress was discovered to be black and blue, but the surprising photo continues to be a source of exploration for scientists of vision perception [Journal of Vision Special Collection, 2016]. For this topics, we collected tweets that contain relevant hashtags from the Garden Hose API. We used four popular hashtags as seeds, #DRESSGATE, #THEDRESS, #WHITEANDGOLD, and #BLACKANDBLUE, then extracted the frequent hashtags from the collected tweets and manually verified those relevant to "The Dress" (Table 3.4). We then generated two groups of hashtags, each of which represents one of two stances: seeing the dress as white and gold, and seeing the dress as black and blue. Among the hashtags in the collected tweets, we extracted one set of hashtags that contain both "black" and "blue", and the other set that contain "white" and "gold". We

[4]http://ciir.cs.umass.edu/irdemo/contention/dataset/

also extracted comparable hashtags in multiple languages, using a list of those color names translated into 80 different languages. We retrieved hashtags that contained the translated words for both "black" and "blue" or both "white" and "gold" in the same language, such as #NEGROYAZUL[5].

The *Brexit referendum,* officially known as the United Kingdom European Union membership referendum, was a referendum that took place on June 23, 2016 in which 51.9% of UK voters voted to leave the EU. While not legally binding, the referendum had immediate political and financial consequences, including the worst one-day drop in the worldwide stock market in history to that date, and the resignation of then-Prime Minister David Cameron.

For Brexit, we downloaded a set of tweet ids[6] collected and released by Milajevs using the tool Poultry [Milajevs and Bouma, 2013]. The data set contained tweet ids related to Brexit from March 7 to August 24th. For each tweet id, we retrieve the corresponding tweet via Twitter Garden Hose API, which allows us to collect 10% random sample of actual tweets if it is included in the sample. Through this process, we were able to obtain 1,222,313 tweets, which is 5.2% of the released Tweet ids for Brexit. Then we used manually curated hashtags to find two stance groups of the tweets, if any stance is revealed in the tweet (Table 3.4).

The *2016 U.S. Presidential Elections* were widely considered one of the most rancorous elections in recent U.S. history, and attracted not only U.S. but also worldwide attention. The two major conflicting stances were with regards to the two main presidential candidates, Donald Trump and Hillary Clinton. To observe the contention trend before, during, and after the voting day, we collected tweets that contain election-related hashtags from Sep 20, 2016 to Nov 30, 2016. We start from the straightforward topic hashtags such as {#election2016, #presidentialelection, #hillaryclinton, #donaldtrump} and a few keywords

[5] 'Negro' means 'black' in Spanish and 'azul' means 'blue'.

[6] See http://www.eecs.qmul.ac.uk/ dm303/brexit

Table 3.2: Twitter Data set with implicit stances

Topic	# Tweets	# Users	Dates
"The Dress"	408.1K	296.9K	Feb. 26-Mar. 9, 2015
Brexit Referendum	1.2M	604.1K	May. 7-Aug. 24, 2016
U.S. Elections	87.4M	10.1M	Sep. 20- Nov. 30, 2016
Rio Olympics	4.6M	1.9M	Aug. 1-Aug.30, 2016
Pokemon Go	3.2M	1.5M	Aug. 1-Aug.30,2016
Nepal Earthquake	49.8K	36.3K	Apr.24-Apr.30,2015
Total	96.9M	14.4M	

such as {president, election, hillary clinton, donald trump} as seeds. Tweets are collected if they contain any of the predefined topic hashtags or keywords. From the collected tweets, we look at the top 50 frequent hashtags and extend the seed hashtag set by adding other relevant hashtags.

To detect the stances, we extracted the top 50 frequent hashtags from the collection. Three expert annotators annotated whether a given hashtag explicitly indicates a stance on which presidential candidate the tweet supports, and we selected only hashtags that all annotators agreed on. Some hashtags contain stances to some extent, but the stances can be either way depending on the context such as #HILLARYBECAUSE and #DRAINTHESWAMP. To take a high-precision, rather than low-recall approach, we extract the set of stance hashtags that three annotators agreed on (Table 4).

In all three cases, we use a high-precision, low-recall approach to detect stances by only assigning a stance to tweets that use an explicit stance hashtag, such as #BLACKANDBLUE or #LEAVEEU. We release a complete list of hashtags used on our website, along with the tweet ids for the collection. While we are certain to miss a large portion of stance-taking tweets that do not use these hashtags, this allows us to be reasonably confident that the stances detected are accurate, which is most useful for the purposes of model validation. We leave analysis of the remaining tweets and other hashtags for future work in stance extraction.

Using the stance hashtags we created, we compute the size of the two stance groups per topic by counting the number of tweets that contain any hashtag from each stance. As an estimation of G_0 (the group with no stance) on each topic, we used all other tweets collected via the Twitter Garden Hose API that day. Specifically, $|G_0|$ = count of all tweets collected $-|G_1| - |G_2|$.

3.4.2.1 Non-controversial topics

In order to validate our model on a range of topics, we also collected Twitter data for three prominent and essentially non-controversial topics: The mobile game Pokemon Go, The 2016 Rio Olympics, and the 2015 Nepal Earthquake. For each of these topics, we examined the top 30 frequent hashtags to check if there exists any conflicting stance. We did not expect to find any conflicting stances in these hashtags, and a close examination of the top 30 hashtags confirmed this. We therefore omit further analysis of these topics for this chapter.

3.4.3 Voting data for Brexit and U.S. Elections

We collected actual voting data for Brexit and the 2016 U.S. Elections. The Brexit voting data, including turnout figures, was released by the UK Electoral Commission [The Electoral Comission, 2016], and was split by Unitary Districts. The EU referendum only had two options, "Remain" or "Leave", which represent two conflicting stances. We considered any non-voters or rejected ballots as having no stance.

For the U.S. Elections, the Federal Election Committee has not released its official results by the time of writing. Nonetheless, we were able to collect the election results from two sources. We used the Popular Vote Tracker [Wasserman, 2017] for certified state results on the 2 major candidates, Donald Trump and Hillary Clinton. Additionally, we used results tabulated on Wikipedia [Wikipedia, 2017], which at the time of writing were official in all but 2 states; these figures included a break-down of results for the three main third party candidates (Johnson, Stein and McMullin) and "Other". Estimated turnout

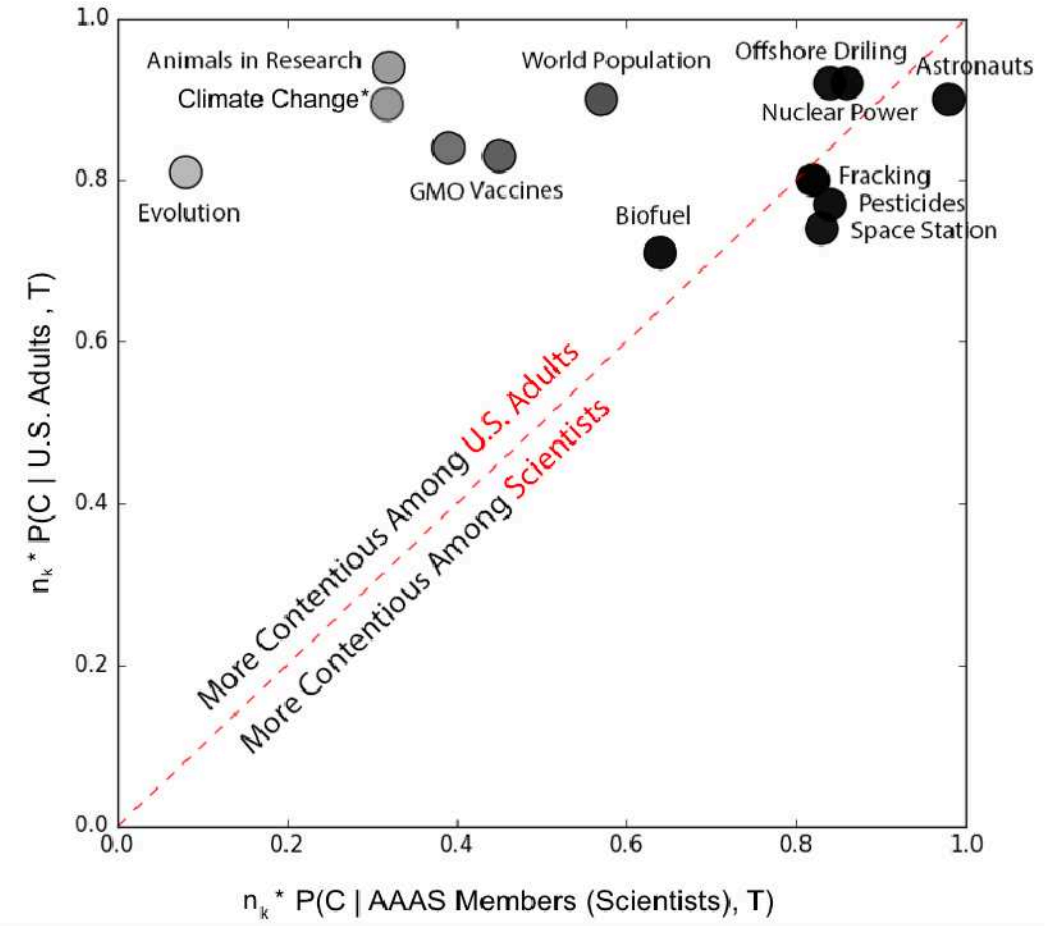

Figure 3.2: Normalized contention in the scientific community vs. general population for several controversial topics. The x=y line represents equal contention among both populations, with dots shaded according to their distance from the line. Note that the Climate Change question had 3 explicit stances, all other questions had 2.

figures were collected from the Elections Project [McDonald, 2017]; we used the reported VEP Highest Office turnout metric, which is available for all U.S. states, to estimate the amount of people holding no stance.

3.5 Model Evaluation

In order to ground our model and ensure that it aligns with actual controversies, we use our model to measure contention on the data sets described above.

Table 3.3: Examples of questions for the topics in Figure 3.2 [Pew Research Center, 2015a, Pew Research Center, 2015b] (bold keywords match point labels).

Issues
Q: Opinion on the increased use of **fracking**: **A:** {Favor, Oppose}
Q: The **space station** has been ... for the country: **A:** {Good investment, Not a good investment}
Q: Thinking about childhood diseases, such as measles, mumps, rubella and polio, do you think... (**label: "vaccines"**) **A:** {All children should be required to be vaccinated, Parents should be able to decide NOT to vaccinate their children}
Q: Do you think it is generally ... to eat foods grown with **pesticides**. **A:** {Safe, Unsafe}

3.5.1 Contention in Polling

We first validate our model using the polls in Table 3.1). We describe a few patterns that emerge.

3.5.1.1 U.S. Scientists vs. General Population

Using the Pew Research data sets (Pew Adults and Pew AAAS in Table 3.1), we are able to examine attitudes towards a number of scientific issues among two populations: U.S. adults and U.S. scientists.

As seen in Figure 3.2, for some topics such as offshore drilling, hydraulic fracturing (fracking), and biofuel, contention was similar between U.S. adults and scientists. On other topics, such as evolution, climate change, and the use of animals in research, contention varied widely depending on the population: the scientific community had low contention for these topics, whereas they were highly contentious among U.S. adults. This result precisely matches prior work's intuitive notion of politically, but not scientifically, controversial topics [Wilson and Likens, 2015]. The graph clearly demonstrates the notion that "there is no controversy" (among scientists) alongside the controversy in general population, with evolution as the most extreme case presented in this data set (98% of AAAS members sur-

Table 3.4: Example hashtags used to identify two stance groups on "The Dress", Brexit and the U.S. Elections. Full list at `http://ciir.cs.umass.edu/irdemo/contention/`.

Topic	Stances	Example Hashtags	# of hashtags
The Dress	Blue and Black	#blackandblue, #notwhiteandgold, #blackandbluedress,#negroyazul ...	49
	White and Gold	#whiteandgold, #whiteandgoldteam, #thedressiswhiteandgold,#blancodorado ...	37
Brexit	Leave EU	#voteleave, #leave, #leaveeu, #betteroffout	4
	Remain EU	#remain, #strongerin, #voteremain, #regrexit, #remainineu	5
U.S. Election	Hillary Clinton	#imwithher, #strongertogether, #dumptrump, #notmypresident ...	10
	Donald Trump	#maga, #trumppence, #trumptrain ...	26

veyed said that "humans and other living things have evolved over time", whereas 31% of the U.S. adults said that they have "existed in their present form since beginning of time").

Interestingly, the food safety of pesticides is equally contentious in both populations, though the direction of contention is exactly the opposite: 68% of AAAS members surveyed said it was safe to eat food grown with pesticides, while 69% of U.S. adults surveyed said it was not. The survey also contained further subdivisions of the AAAS populations surveyed (e.g. based on their degree earned, employment status, and area of expertise) [Pew Research Center, 2015a, Pew Research Center, 2015b], though results were largely similar to the general AAAS population, so we omit them here.

3.5.1.2 Contention over time for "hot button" topics

The Gallup data set gives us access to changing contention over time for several controversial topics in the U.S. We selected three topics: the death penalty for murder, legalization of marijuana, and legalization of same-sex marriage. As seen in Figure 3.3, clear trends emerge when contention is mapped over time. For example, marijuana legalization had consistently low contention in the early '70s (when less than 20% of the population thought it should be legalized); support for the death penalty was high (and contention low) during the '90s. Interestingly, contention for both marijuana legalization and same-sex marriage peaked recently, and is now going down as the support for each of these has crossed the threshold of 50% around 2012. For the death penalty, contention between sub-populations

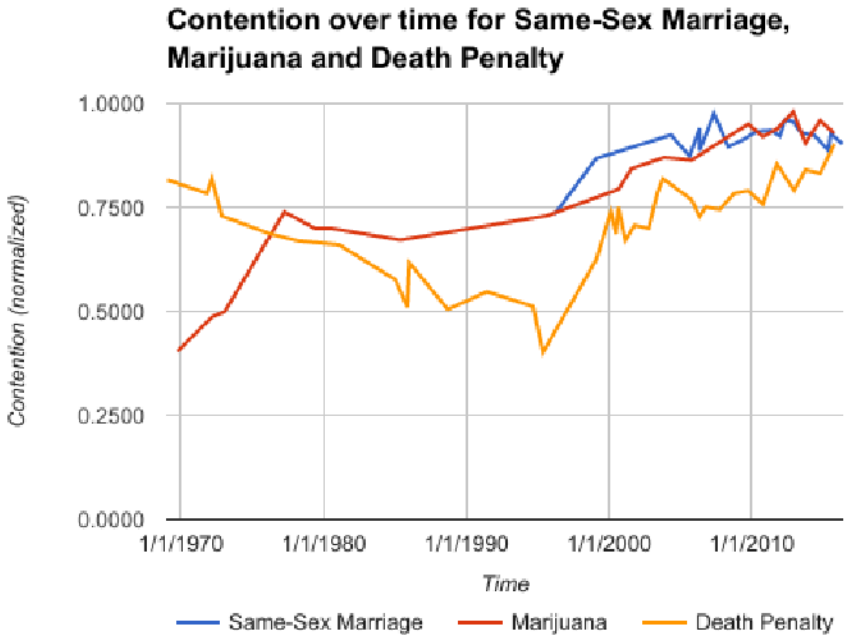

Figure 3.3: Contention over time for three controversial topics (normalized for two stances). Contention around 1.0 masks the trend direction of the stances, e.g. in the case of growing approval for same-sex marriage and marijuana in recent years. Results for "Death Penalty" prior to 1969 are omitted.

in the U.S. varied widely; for example, contention was higher among the black and Hispanic populations, and higher for democrats (full results omitted for space considerations).

3.5.1.3 Per-state distribution of Contention in the United States

Using the iSideWith data set, we measured normalized contention nation-wide and per-state on each of the 52 topics available. The two least contentious questions nation-wide were "Should National Parks continue to be preserved and protected by the federal government?" ($n_k * P(c|US,t)$ = 0.26), and "Should every person purchasing a gun be required to pass a criminal and public safety background check?" ($n_k * P(c|US,t)$ = 0.39). Several topics had over 0.99 normalized contention nation-wide, such as "Should the U.S. formally declare war on ISIS?" and "Would you support increasing taxes on the rich in order to reduce interest rates for student loans?", among others. We present the

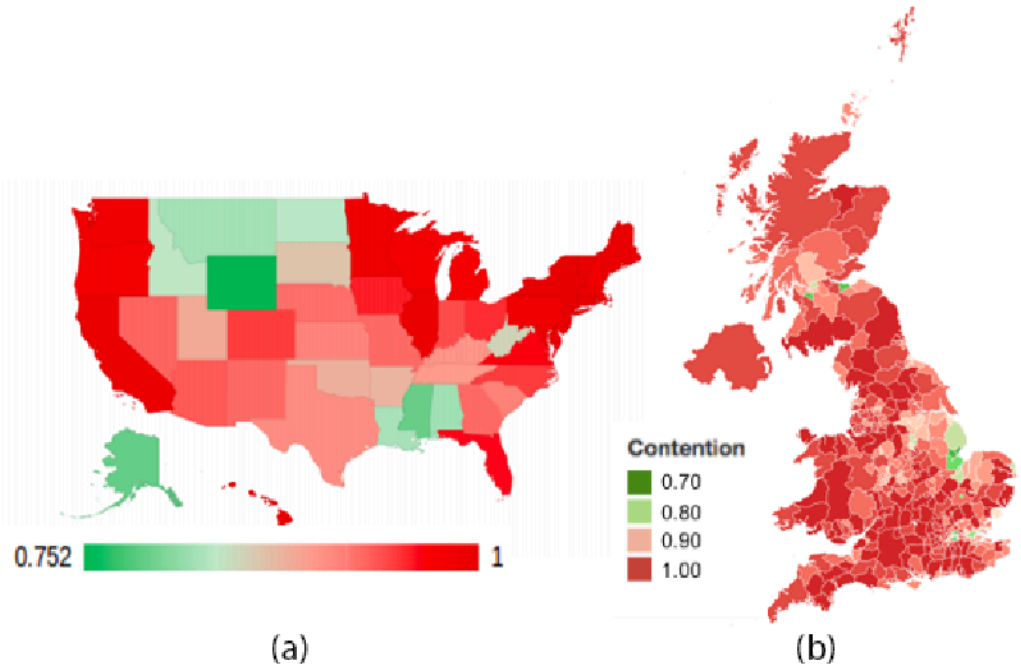

Figure 3.4: (a) Per-state contention for "Do you support increased gun control?" (normalized for two stances). (b) Contention by voting district in the UK (normalized for two stances) [The Electoral Commission, 2016]. Interactive maps for all iSideWith issues are available at http://ciir.cs.umass.edu/irdemo/contention/isidewith/ .

per-state contention for one such topic in Figure 3.4, which shows how contention varies geographically. An interactive demo with per-state contention on all 52 topics is available at http://ciir.cs.umass.edu/irdemo/contention/isidewith/.

3.5.2 Contention on Twitter

From the Twitter data collected above, we report contention for our three controversial topics: "The Dress", Brexit and the U.S. elections. For each topic, we calculate two types of daily contention trends: one, only among the tweets exhibiting a stance on the topic on that day, and the other among all of the Twitter posts on that day, i.e., including G_0. A visible pattern emerges, where contention only among the population that exhibits a stance is consistently high throughout, whereas including G_0 shows marked peaks of contention around notable event times. For example, in the U.S. Elections case, small peaks appear

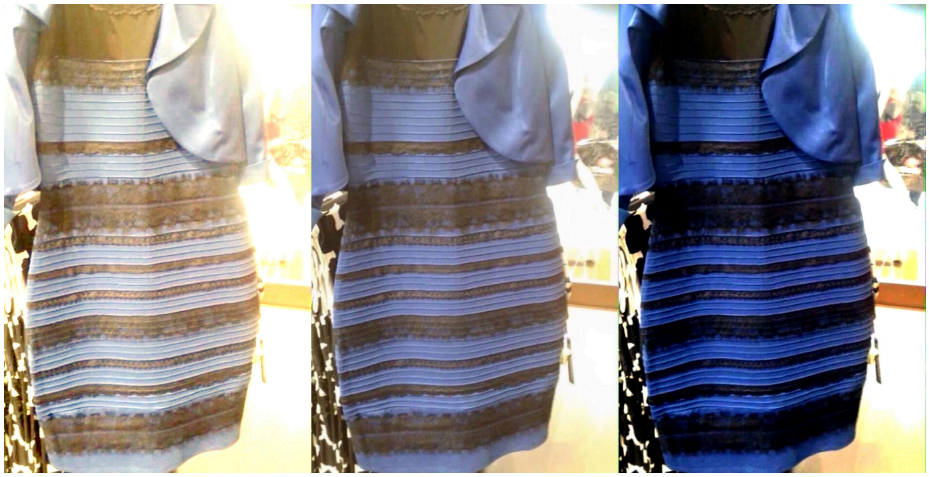

Figure 3.5: "The Dress" photo, which went viral after people strongly disagreed on its colors. The original photo is in the center. Image credit: Wired [Rogers, 2015].

on the days of the presidential debates, and upon release of the extremely controversial Hollywood Access tape, with a much larger peak on election day. This showcases the strength of our model and its ability to track the difference between contention among the group for which the topic is salient ($G_1 \cup G_2$), as opposed to the entire population.

Comparison to external sources. We compare $n_k * P(c|G_1 \cup G_2, T)$ from Twitter across a series of dates, with that calculated from external sources: the Buzzfeed poll on "The Dress" ($n_k * P(c|G_1 \cup G_2, T)$ = 0.88) [Holderness, 2015], voting results on Brexit ($n_k * P(c|G_1 \cup G_2, T)$ = 1.00) [The Electoral Comission, 2016], and the popular vote in the U.S. Elections measured for the two main candidates ($n_k * P(c|G_1 \cup G_2, T)$ = 0.89). Additionally, Figure 3.4(b) shows the voting contention for each Unitary District of the UK (local Ireland results were not available), demonstrating the geographical variance of contention. Gibraltar, an extreme outlier both geographically and contention-wise, is omitted from the map ($n_k * P(c|Gibraltar, Brexit) = 0.16$). The extremely low contention makes sense: Gibraltar is geographically located inside Europe, and 95.9% of its voters voted "Remain".

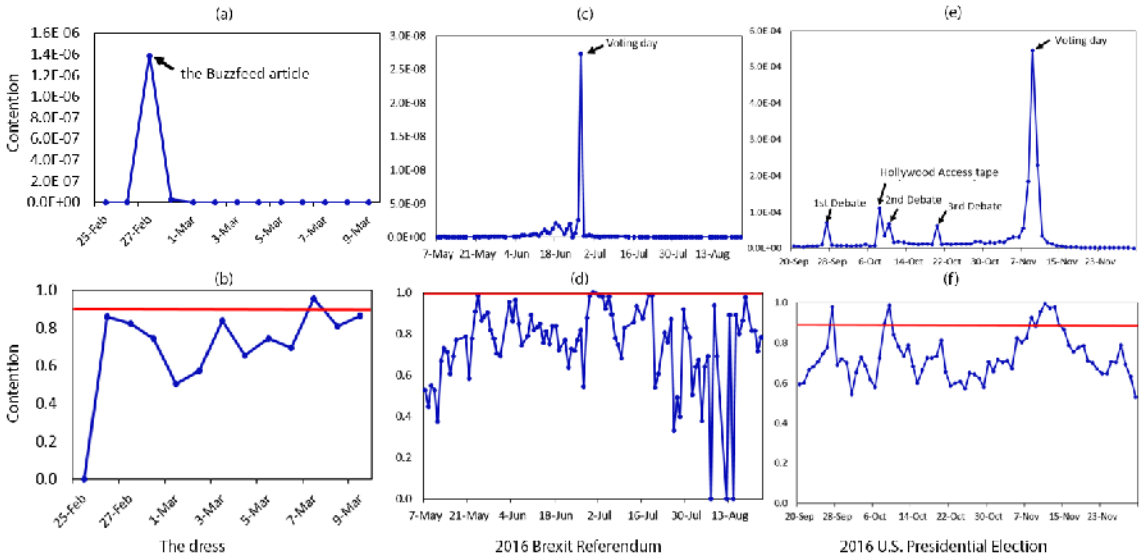

Figure 3.6: Normalized contention among all daily tweets by date for "The Dress" (left), Brexit (center) and 2016 U.S. Elections (right), reported among all Gardenhose tweets that day (top) or only among those with an explicit stance (bottom). Notable peaks are annotated with associated events around that time. All dates are in UTC. The horizontal lines in (b), (d), (f) show the normalized contention from alternate sources ("The Dress", 0.88; Brexit, 1.00; U.S. Elections, 0.89).

Turnout in voting. For the 2016 United States elections and Brexit, we measured contention with or without estimated turnout figures. In both cases, G_0 was set as the number of eligible voters (official in the UK, estimated in the U.S.) who did not vote. Contention decreases markedly when voter turnout is factored into the model. For the extremely divisive U.S. elections, normalized contention dropped from 0.89 to 0.31 when factoring in the estimated 41.1% of eligible voters that did not go to the ballots on election day. A similar pattern is observed for Brexit.

Contention and Third-Party Votes. We briefly analyzed the results of contention in the U.S. Elections as measured on the two main candidates as well as the three main third-party candidates, Johnson, Stein and McMullin and a sixth category reported as "Other" [Wikipedia, 2017]. This yields a total of 6 stances, and a few interesting patterns are revealed when examining this six-way contention. For example, measured only on Trump

and Clinton, normalized contention is nearly the lowest in Utah, but is highest of all states when considering the third party candidates. This makes sense when considering that Evan McMullin received 21.3% of the vote in that state.

3.6 Discussion

Our population-based contention model offers a new way of quantifying controversies, and a new way to understand multiple observed phenomena, only some of which we explicitly covered in this chapter. For example, a conflict between two populations will often have low contention internally since each is fairly consistent with a specific stance, but when the two populations are observed together, the combination is highly contentious. Small, community-specific controversies can now be quantified as well; a certain topic might be extremely controversial in a tight-knit population, while the rest of the world is starkly in G_0, either oblivious or apathetic to the controversy. Other population-dependent contention levels can be observed elsewhere, for example in the case of racial tensions around police brutality in the U.S. As demonstrated in Figure 3.2, we can use this model to quantify the aforementioned high-stakes public opinion controversies over scientifically well-understood phenomena. In the scientific community, topics such as climate change, evolution and vaccines are in consensus, while in the general U.S. population, their contention remains high.

For the purpose of model validation, we intentionally chose to use a high-precision, low-recall manual curation process to classify stances. However, we note that this high-quality curation is not central to the contention model: implicit or inferred stances can be used in the same manner. In fact, this stance detection process can be automated, as demonstrated by recent work [Coletto et al., 2016, Garimella et al., 2016], and such advances are synergistic with our contention metric.

3.6.1 Model Limitations

As noted in Section 3.3, our model allows for overlapping stances, which are in practice very challenging to estimate. The added constraints of mutually exclusive stances which all conflict equally, make the model extremely practical and easy to estimate; however, one must take care to ensure that the stances fed to the model are indeed mutually exclusive, otherwise the conclusions may not hold. The constraints are certainly true for many controversial topics, but not all of them. For example, for "The Dress" we know there was a subset of people who in practice saw both color combinations, which we did not take into account. Even for mutually-exclusive stances, comparison between issues with a varying number of stances may be complicated by the normalization factor, and further exploration is needed to understand this effect better. Additionally, if multiple stances lie on a spectrum between two extremes, it does not make sense to consider them all equally conflicting. In such a case, recasting the *holds* and *conflicts* functions to return a real value in the [0,1] range instead of a binary value may be a better fit; such a "variable edit distance" function is well known in the bioinformatics space, and existing work in that space could be leveraged for contention. Such a recasting might result in a more nuanced characterization of multi-stance controversies and allow a better comparison between them and two-sided controversies. We leave these analyses for future work.

3.6.2 Future work

Our theoretical model of contention points the way to several possible avenues of future research. As mentioned above, stance extraction is a growing research topic [Coletto et al., 2016, Garimella et al., 2016], and automated stance extraction can certainly be applied to improve the detection and measurement of contention in the near future. An alternative conception of contention could conceivably start from groups rather than individuals, in a model which would explain stance as a conclusion of group membership [Kahan, 2015]. The differentiation between overlapping and mutually exclusive stances might be useful for

classification of controversiality, reminiscent of a recent partitioning approach to measuring controversy [Garimella et al., 2016]. Likewise, evaluating multiple stances is a challenge we leave for future work due to the sparsity of existing evaluation data.

Our work also clearly calls out the need for more research into the additional dimensions of controversy, beyond contention. For example, "importance" as a dimension of controversy allows for further examination. Conceivably, importance could possibly be automatically extracted for different populations, à la the efforts currently placed on stance extraction; for example, an indicator for importance might be on-the-ground protests in various regions [De Choudhury et al., 2016]. Importance also relates to the recent study of the relationship between controversy and conversation vs. discomfort [Chen and Berger, 2013]; combining their work with our model suggests that high-importance controversies may increase discomfort whereas low-importance controversies may increase conversation, as in the case of "The Dress". Alternative dimensions that might contribute to our hypothesized controversy model, which we have yet to explore, include notions of "conviction" (how likely is a person to change their stance?), "identity-centrality" (how central is this controversy to the individual's identity?), as well as "loudness" or "influence": all people are considered equally when evaluating contention, when in fact the stances of certain "thought leaders" may have a disproportionate impact by increasing the diffusion of their stances.

3.7 Conclusions

This chapter introduced our first set of contributions, those related to the definition of controversy. Drawing on work from a variety of disciplines, we hypothesized a new theoretical model for re-conceptualizing controversy (contribution 1.3.1.1). We redefine controversy as population-dependent, and as multi-dimensional rather than a single quantity. We posited that contention is one such dimension, and presented preliminary evidence that importance is another possible dimension. Our contention measure and the hypoth-

esized controversy model hold significant promise in offering a deeper understanding of the nature of controversies, increasing the likelihood of reproducibility of future work, and holding implications for social science, humanities and computer science research on controversies, with civic, social and science-communication implications. We leave further exploration of this multi-dimensional controversy conception to future work.

We then proposed a new measure, contention, which mathematically quantifies the notion of "the proportion of people disagreeing on this topic" in a population-dependent fashion (contribution 1.3.1.2). Our framework departed from most existing work about controversy in a two major ways. First, in contrast to prior work which considers controversy to have a single global value, we define contention not only in terms of its topic, but also in terms of the population being observed. Second, our model accounts for participants in the population who hold no stance with regards to a specific topic, and also allows for multiple stances rather than just two opinions. We validated our theoretical model on a wide variety of data sets from both off- and online sources, ranging from large informal online polls and Twitter data, through statistically calibrated phone surveys, and actual voting records.

The novel framework we introduced in this chapter allows us to quantify a wide variety of phenomena, such as the difference between scientific controversies and political ones, the change in contention over time, and local or cultural patterns in contention (contribution 1.3.1.3). We used our diverse sources to demonstrate that the contention measure holds explanatory power for a wide variety of observed phenomena that cannot be explained under previous global controversy views:

1. Controversies over climate change, vaccines, and other topics that are well within scientific consensus, and which scientists often say "there is no controversy". These can be explained under the new model: there is indeed no controversy within the scientific community, while there is still controversy among the general population in certain regions.

2. International conflict (such as the Israeli-Palestinian conflict) can be understood as exhibiting high contention at the global level, often with moderate to low contention within each participating nation.

3. Well-documented polling variations in controversy among certain populations or interest groups, such as different attitudes toward corporal punishment among different racial groups, can be easily modeled under population-dependent contention.

4. Topics that are controversial only in certain geographical regions or among certain interest groups can likewise be modeled.

As a side effect of our work on contention, we created a Twitter data set of nearly 100 million tweets, for several popular topics in the last eighteen months, including three prominent controversies (the 2016 U.S. Elections, the UK referendum on leaving the EU, commonly known as Brexit, and "The Dress", a photo that went viral when people disagreed on its colors). We publicly release this rich data set and make it available for the benefit of the research community (contribution 1.3.1.4).

In Chapter 6, we will add an additional validation step for contention by creating a derivation of our contention model that can be applied to Wikipedia, and evaluate it on 2000 Wikipedia articles. We will compare our probabilistic, theoretically-motivated model to the M measure, a previous work that is heuristic in nature [Sumi et al., 2011].

CHAPTER 4

AUTOMATED CONTROVERSY DETECTION ON THE WEB

In this chapter, we introduce the problem of detecting controversial topics on the web, and describe an algorithm to solve this problem. We first demonstrate an oracle-based approach as a proof of concept that the problem could be solved by connecting web pages to related Wikipedia articles, thus leveraging the rich metadata available in Wikipedia. We then introduce a fully-automated, weakly-supervised approach to detect controversial topics on arbitrary web pages, thus offering the first fully-automated solution to the problem of detecting web pages on controversial topics. We use automatically generated scores for Wikipedia, including M, which is related to our contention measure from Chapter 3. We consider our system as distantly-supervised [Riedel et al., 2010] since we use heuristic labels for neighboring Wikipedia articles, in addition to a smaller amount of truth data on the web. Much of the work in this chapter was previously published elsewhere [Dori-Hacohen and Allan, 2013, Dori-Hacohen and Allan, 2015].

4.1 The problem of controversy detection on the web

In our early work [Dori-Hacohen and Allan, 2013], we posed a new problem in the Information Retrieval community: does a webpage represent a controversial topic? To the best of our knowledge, this problem had not been formulated as such before, though several special cases have been explored by previous researchers (e.g. in Wikipedia, where the rich metadata offers additional signals).

As mentioned in Chapter 1, we are interested in techniques that encourage and facilitate healthy debates, allowing users to critically approach these issues. One way to do so is to

alert users when their search results represent a perspective on a controversial issue; for example, imagine a warning presented at the top of a web page: "This webpage represents one of several perspectives on a controversial topic." To do so, we need to answer a non-trivial question: "Is this topic controversial?" [1]

Note that our goal differs from "diversifying" search results, wherein – perhaps – each of the perspectives might be presented in a ranked list. Instead, we aim to identify whether a single page *in isolation* discusses a topic with divergent stances, and thus, controversy.

We approach this as an estimation problem: determining the level of controversy in a topic, while thresholding it for binary classification. We utilize a supervised k-nearest-neighbor classifier on web pages that uses labeled estimates of controversy in Wikipedia articles to determine the likelihood that a web page is controversial itself. Essentially, a page similar to controversial pages is likely controversial itself. Our choice of Wikipedia articles as labeled neighbors is motivated both by topical coverage as well the possibility of using unsupervised labels of controversy from prior work.

As part of this work, we create and release a new data set, described below. Our set is the first collection of its kind, which includes 377 web pages that were manually judged as controversial or not, as well as over 1700 Wikipedia articles on related topics. Hypothesizing that controversial material is often highly opinionated, we compare our results to a sentiment analysis classifier.

4.2 Controversy Annotation Data Set

To investigate the feasibility of our approach, we construct and release a suitable data set. [2] We hypothesize that we can detect controversy indirectly by using the controversial-

[1] Crucially, answering this question does not entail passing moral judgment on the web pages.

[2] This data set is freely available at http://ciir.cs.umass.edu/downloads

Table 4.1: Data set size and annotations. "NNT" denotes the subset of Wikipedia articles that are Nearest Neighbors of the webpages Training set.

Webpages			
Set	Seeds	Pages	Controversial
Training	Wikipedia	248	74 (29.8%)
Testing	Wikipedia	129	49 (38.0%)

Wikipedia articles			
Set	Articles	Annotated	Controversial
All	8,755	1,761	282 (16.0%)
NNT	4,060	853	115 (13.5%)

ity of Wikipedia articles that are similar to the starting webpage. Thus, our data set also includes judgments on the controversiality of Wikipedia articles.

Our data set, described in Table 4.1, was created as follows. We selected 41 seed articles from Wikipedia. The articles were chosen based on their implied level of controversy, with some clearly controversial ("Abortion") and others clearly not controversial ("Mary Poppins"). We used only the Wikipedia article's title as a query to the blekko search engine[3]. From up to top 100 results returned for queries, we selected only webpages that also appeared in ClueWeb09 category B[4] to allow reproducibility. We also omitted Wikipedia articles, pages that could not be displayed properly, and pages that had no nearest neighbors among the Wikipedia articles (see below and Section 4.3.1), leaving 377 web pages over the 41 seed topics.

We split this collection into training and testing sets based on the seeds – since our pages were not chosen independently. We wanted approximately a 60-40 split, so we divided our seeds randomly into 30% whose "related" webpages were labeled as all training, 20% as all testing, and 50% of the seeds whose webpages were split, as one group, at a 60-40 ratio between the training and testing collections. The final distribution of the collections

[3]http://blekko.com

[4]http://lemurproject.org/clueweb09/

Table 4.2: Inter-annotator agreement. *Results are shown separately for 2 and 3 annotators that rated the same page.*

All (2 or 3)	Pages		Judgments	
Total	344		851	
Agreement	224	(65.1%)	551	(64.7%)
Disagreement (all)	120	(34.9%)	300	(35.3%)
2 Annotators	**Pages**		**Judgments**	
Total	181		362	
Agreement	121	(66.9%)	242	(66.9%)
Disagreement (Tie)	60	(33.1%)	120	(33.1%)
3 Annotators	**Pages**		**Judgments**	
Total	163		489	
Agreement	103	(63.2%)	309	(63.2%)
2-1 Disagreement	60	(36.8%)	180	(36.8%)

differed slightly due to our selection method, as shown in Table 4.1: the training set had a lower proportion of controversial pages than the testing set (29.8% vs. 38.0%).

We created an annotation tool to capture the controversy level of these pages. We ask how controversial is the topic discussed by the webpage, and the options were: "1 - clearly controversial", "2 - possibly controversial", "3 - possibly non-controversial", or "4 - clearly non-controversial". By design, 344 of the 377 pages were annotated by more than one annotator for 851 total judgments. Table 4.2 summarizes the agreement among the annotators. 65.1% of the pages had complete agreement, accounting for 64.7% of the judgments. Another 17.4% had a majority (2 of 3) vote, with 17.4% of the pages tied among two annotators.

Our oracle-based approach also relies on labeled data from Wikipedia. We used a variation of the annotation tool to judge the controversiality of Wikipedia articles. For each of the 377 pages we found its nearest Wikipedia articles using queries to blekko (as described in Section 4.3.1), for a total of 8755 unique Wikipedia articles. We annotated as many top-ranking Wikipedia articles as we could, resulting in 1761 Wikipedia articles judged by our annotators, as shown in Table 4.1. Of these, 331 were annotated by more

than one annotator, and they agreed on 81.6% of the Wikipedia pages. (Of the Wikipedia articles annotated in the set, 4,060 were the Nearest Neighbors associated with the Training set ("NNT" in Table 4.1), which we use later (see Section 4.4.4). For evaluation, we use Precision, Recall, Accuracy, F_1 and $F_{0.5}$ using the classic IR sense of these metrics, with "controversial" and "non-controversial" standing in for "relevant" and "non relevant", respectively.) This data set was also used in Section 5 in improving automated controversy detection within Wikipedia.

Whenever a webpage or Wikipedia article was annotated more than once, we took the average value of all the judgments (in the range [1..4]) as its controversy score, which we use in our approach and evaluation. To convert into a binary value, any score below a threshold of 2.5 (the midpoint of our 4-point range) is considered controversial.

Additionally, the data set contains 3,430 annotations of query-article combinations (see Section 4.4.1 for a description).

4.3 Nearest Neighbor approach

Our approach to detecting controversy on the web is a nearest neighbor classifier that maps webpages to the Wikipedia articles related to them. We start from a webpage and find Wikipedia articles that discuss the same topic; if the Wikipedia articles are controversial, it is reasonable to assume the webpage is controversial as well. The choice to map specifically to Wikipedia rather than to any webpages was driven by the availability of the rich metadata and edit history on Wikipedia, as discussed in prior sections.

In our own work, we first demonstrate that this approach works using human judgment as an oracle, in a supervised manner. We later extend this work to use automatically generated labels. We consider this second approach as a distantly-supervised classifier in the relaxed sense (c.f. [Riedel et al., 2010]), since we are using automatically-generated labels, rather than truth labels, for an external data set (Wikipedia) rather than the one we are training on (web). While some of these labels were learned using a supervised classifier on

Wikipedia, none of them were trained for the task at hand, namely classifying webpages' controversy.

To implement our nearest neighbor classifier, we use several modules: matching via query generation, scoring the Wikipedia articles, aggregation, thresholding and voting.

4.3.1 Matching via Query Generation

We use a query generation approach to map from webpages to the related Wikipedia articles. The top ten most frequent terms on the webpage, excluding stop words, are extracted from the webpage, and then used as a keyword query restricted to the Wikipedia domain and run on a commercial search engine. We use one of two different stop sets, a 418 word set (which we refer to as "Full" Stopping [Callan et al., 1992]) or a 35 word set ("Light" Stopping [Manning et al., 2008]). Wikipedia redirects were followed wherever applicable in order to ensure we reached the full Wikipedia article with its associated metadata; any talk or user pages were ignored.

We considered the articles returned from the query as the webpage's "neighbors", which will be evaluated for their controversy level. Based on the assumption that higher ranked articles might be more relevant, but provide less coverage, we varied the number of neighbors in our experiments from 1 to 20, or used all articles containing all ten terms. A brief evaluation of the query generation approach is presented in Section 4.4.1.

4.3.2 Wikipedia labels (Automatically-generated and human)

The Wikipedia articles, found as neighbors to webpages, were labeled with several scores measuring their controversy level. In our supervised work, we used human judgments as an "oracle" for the controversy score. For our weakly-supervised model, we use three different types of automated scores for controversy in Wikipedia, which we refer to as **D**, **C**, and **M** scores. All three scores are automatically generated based on information available in the Wikipedia page and its associated metadata, talk page and revision history.

While we use a supervised threshold on the scores, the resulting score and prediction can be generated with no human involvement.

- Oracle Scores: as described in Section 4.2, we have labels of controversy on 20% of these articles (with preference towards articles ranking higher in the retrieval). We use our annotators' judgments of Wikipedia articles whenever they are available. We aggregate the score over k neighbors of the webpage to receive a final controversy score.

- The D score: a binary score that tests for the presence of **D**ispute tags that are added to the talk pages of Wikipedia articles by its contributors [Kittur et al., 2007, Sepehri Rad and Barbosa, 2012]. These tags are sparse and therefore difficult to rely on [Sepehri Rad and Barbosa, 2012], though potentially valuable when they are present. We test for the presence of such tags, and use the results as a binary score (1 if the tag exists or -1 if it doesn't). Unfortunately, the number of dispute tags available is very low: in a recent Wikipedia dump, only 0.03% of the articles had a dispute tag on their talk page. This is an even smaller data set than the human annotations we collected; the overlap between these articles and the 8,755 articles in our data set is a mere 165 articles.

- The C score: a metadata-based regression that predicts the controversy level of the Wikipedia article using a variety of metadata features (e.g. length of the page and its associated talk page, number of editors and of anonymous editors). This regression is based on the approach first described by Kittur et al. [Kittur et al., 2007]. We use the version of this regression as implemented and trained recently by Das et al. [Das et al., 2013], generating a floating point score in the range (0,1).

- The M score: as defined by Sumi, Yasseri and their colleagues, is a different way of estimating the controversy level of a Wikipedia article, based on the concept of mutual reverts and edit wars in Wikipedia [Sumi et al., 2011]. As we discussed in

Section 6.1, this score is effectively measuring the dimension of controversy that we defined as "contention" in Chapter 3. Their approach is based on the number and reputation of the users involved in reverting each others' edits, and assumes that "the larger the armies, the larger the war" [Yasseri et al., 2012]. The score is a positive real number, theoretically unbounded (in practice it ranges from 0 to several billion).

4.3.3 Aggregation and Thresholding

In both models, the score for a webpage is computed by taking either the maximum or the average of all its Wikipedia neighbors' scores, a parameter we vary in our experiments.

In our fully-supervised model, the oracle score is the only score available.

However, in the weakly-supervised model, an additional thresholding step is added. After aggregation, each webpage has 3 "controversy" scores from the three scoring methods (**D**, **C** and **M**). We trained various thresholds for both **C** and **M** (see Section 4.4.4), depending on target measures.

4.3.4 Voting

In the weakly-supervised model, in addition to using each of the three labels in isolation, we can also combine them by voting. We apply one of several voting schemes to the binary classification labels, after the thresholds have been applied. The schemes we use are:

- Majority vote: consider the webpage controversial if at least two out of the three labels are "controversial".
- Logical Or: consider the webpage controversial if any of the three labels is "controversial".
- Logical And: consider the webpage controversial only if all the three labels are "controversial".

- Other logical combinations: we consider results for the combination ($Dispute \vee (C \wedge M)$), based on the premise that if the dispute tag happens to be present, it would be valuable[5].

4.4 Experimental Setup

We use the data set described in Section 4.2. We treat the controversy detection problem as a binary classification problem of assigning labels of "controversial" and "non-controversial" to webpages. For evaluation, we use Precision, Recall, Accuracy, F_1 and $F_{0.5}$ using the classic IR sense of these metrics, with "controversial" and "non-controversial" standing in for "relevant" and "non relevant", respectively. We present a brief evaluation for the query generation approach and our baseline runs before turning to describe our results for the controversy detection problem.

4.4.1 Judgments from Matching

A key step in our approach is selecting which Wikipedia articles to use as nearest neighbors. In order to evaluate how well our query generation approach is mapping webpages to Wikipedia articles, we evaluated the automated queries and the relevance of their results to the original webpage. This allows an intrinsic measure of the effectiveness of this step - independent of its effect on the extrinsic task, which is evaluated using the existing data set's judgments on the webpages' controversy level[6]. We annotated 3,430 of the query-article combinations (out of 7,630 combinations total) that were returned from the search engine; the combinations represented 2,454 unique Wikipedia articles. Our annotators were presented with the webpage and the titles of up to 10 Wikipedia articles in alphabetical order (not ranked); they were not shown the automatically-generated query. The

[5]D's coverage was so low that other voting combinations were essentially identical to the majority voting; we therefore omit them.

[6]As mentioned above, this data set is publicly released - see http://ciir.cs.umass.edu/downloads

annotators were asked to name the single article that best matched the webpage, and were also asked to judge, for each article, whether it was relevant to the original page. Figure 4.1 shows how the ranked list of Wikipedia articles were judged. In the figure, it is clear that the top-ranking article was viewed as highly on topic but then the quality dropped rapidly. However, if both "on-topic" judgments are combined, a large number of highly or slightly relevant articles are being selected. Considering the rank of the best article as the single relevant result, the Mean Reciprocal Rank for the data set was 0.54 (if the best article was "don't know" or "none of the above", its score was zero).

4.4.2 Baselines

As a new problem, no obvious baseline algorithm exists for web classification. However, since controversy can arguably be described as the presence of strong opposing opinions, a natural baseline is a sentiment analysis classifier. For our baseline, we took a sentiment analysis approach based on a logistic regression classifier [Aktolga and Allan, 2013] trained to detect presence of sentiment on the webpage, whether positive or negative; sentiment is used as a proxy for controversy. We add single-class and random baselines (average of three runs). Finally, the best results from our supervised, oracle-based work are reported for comparison.

4.4.3 Parameters for Weakly-Supervised approach

As described in Section 4.3, we varied several parameters in our nearest neighbor approach:

1. Stopping set (Light or Full)

2. Number of neighbors (k=1..20, or no limit)

3. Aggregation method (average or max)

4. Scoring or voting method (C, M, D; Majority, Or, And, $D \vee (C \wedge M)$)

5. Thresholds for C and M (one of five values, as described in Section4.4.4).

Table 4.3: Results on Testing Set. Results are displayed for the best parameters on the training set, using each scoring method, optimized for F_1, Accuracy and $F_{0.5}$. The overall best results of our fully-automated runs, in each metric, are displayed in bold; the best oracle results (rows 12-14) and baseline results (rows 15-19) are also displayed in bold. See text for discussion.

	Parameters							Test Metric				
#	Stop	Score	k	agg	Thres C	Thres M	Target	P	R	F_1	Acc	$F_{0.5}$
1	Full	M	8	avg	–	*84930*	F_1, Acc	0.55	0.67	0.61	0.67	0.57
2	Light	M	8	max	–	*2.85*$\times 10^6$	$F_{0.5}$	0.63	0.63	0.63	**0.72**	**0.63**
3	Light	C	15	max	*0.17*	–	F_1	0.57	0.71	0.64	0.69	0.60
4	Light	C	7	avg	*4.18*$\times 10^{-2}$	–	Acc, $F_{0.5}$	**0.64**	0.57	0.60	0.71	0.62
5	Light	D	19	max	–	–	F_1	0.43	0.57	0.49	0.55	0.45
6	Full	D	5	max	–	–	Acc	0.53	0.37	0.43	0.64	0.49
7	Light	D	6	max	–	–	Acc, $F_{0.5}$	0.44	0.35	0.39	0.58	0.41
8	Light	Maj.	15	max	*0.17*	*2.85*$\times 10^6$	F_1	0.59	0.73	**0.65**	0.70	0.61
9	Full	Maj.	5	max	*4.18*$\times 10^{-2}$	*2.85*$\times 10^6$	Acc, $F_{0.5}$	0.59	0.61	0.60	0.69	0.59
10	Light	And	no	max	*0.17*	*84930*	F_1, Acc, $F_{0.5}$	0.52	0.51	0.51	0.64	0.52
11	Light	D|CM	7	avg	*4.18*$\times 10^{-2}$	*84930*	Acc, $F_{0.5}$	0.63	0.55	0.59	0.70	0.61
12	Oracle-based, best run for P, Acc and $F_{0.5}$							**0.69**	0.51	0.59	**0.73**	**0.65**
13	Oracle-based, best run for R							0.51	**0.84**	0.64	0.64	0.56
14	Oracle-based, best run for F_1							0.60	0.69	**0.64**	0.70	0.61
15	Sentiment							0.38	0.90	0.53	0.40	0.43
16	Random_{50}							**0.42**	0.53	0.47	0.54	**0.44**
17	$\text{Random}_{29.8}$							0.23	0.19	0.21	0.61	0.22
18	All non-controversial							0	0	0	**0.62**	0
19	All Controversial							0.38	**1.00**	**0.55**	0.38	0.43

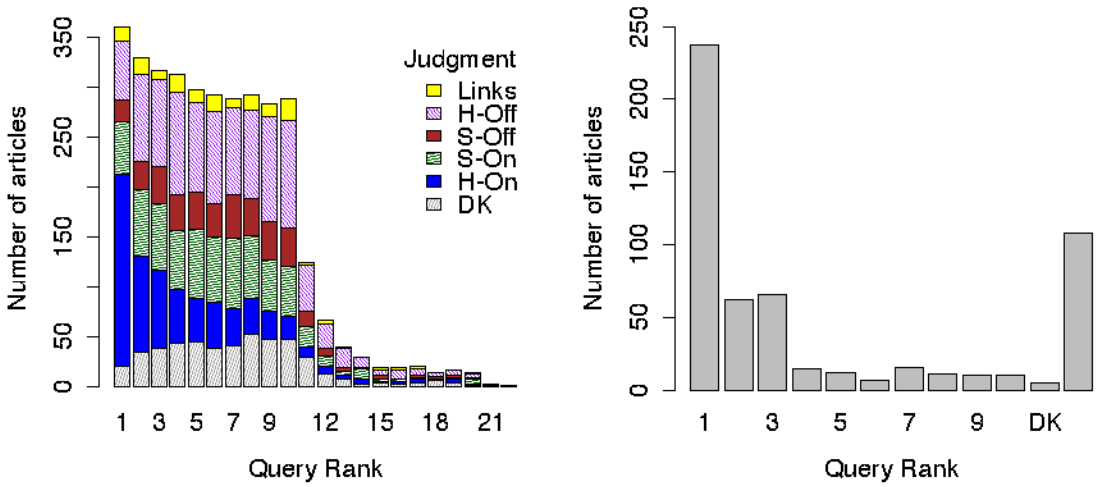

Figure 4.1: Evaluation of Matching scheme. Left: Judgments on Wikipedia articles returned by the automatically-generated queries, by rank. Annotators could choose one of the following options: H-On="Highly on [webpage's] topic", S-On="Slightly on topic", S-Off="Slightly off topic", H-Off="Highly off topic", Links="Links to this topic, but doesn't discuss it directly", DK="Don't Know". Right: Frequency of page selected as best, by rank. DK="Don't Know", N="None of the above".

These parameters were evaluated on the training set and the best runs were selected, optimizing for F_1, $F_{0.5}$ and Accuracy. The parameters that performed best, for each of the scoring/voting methods, were then run on the test set.

4.4.4 Threshold training

C and **M** are both real-valued numbers; in order to generate a binary classification, we must select a threshold above which the page will be considered controversial. (**D** score is already binary.) Since our data set included annotations on some of the Wikipedia articles, we trained the thresholds for **C** and **M** for the subset of articles associated with the training set (labeled "NNT" in Table 4.1). The Precision-Recall curve for both scores is displayed in Figure 4.2. We select five thresholds for the two scoring methods, based on the best results achieved on this subset for our measures.

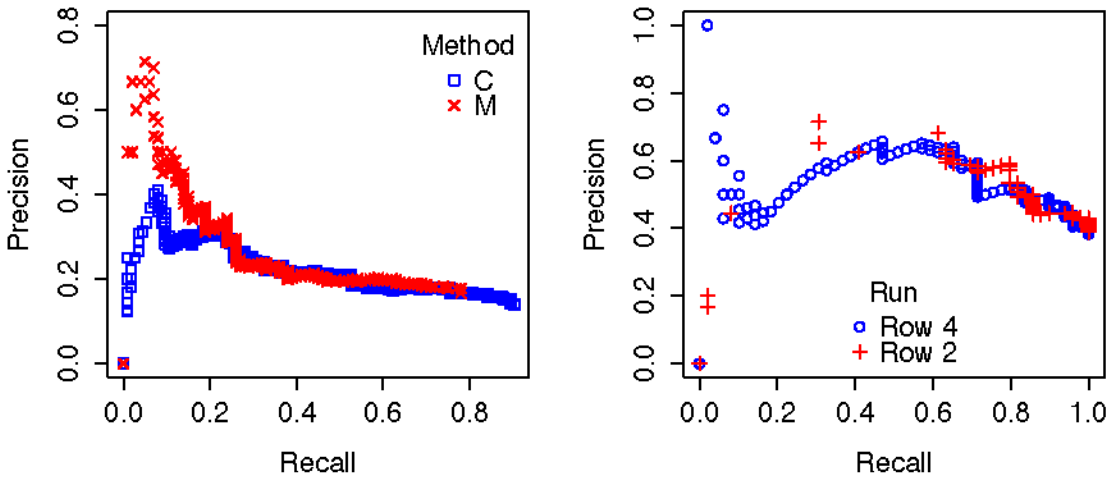

Figure 4.2: Precision-Recall curves (uninterpolated). Left: PR curve for C and M thresholds on the Wikipedia NNT set. Right: PR curve for select runs on the Test set. Row numbers refer to Table 4.3.

For comparison, we also present single-class acceptor baselines on this task of labeling the Wikipedia articles, one which labels all pages as non-controversial and one which labels all pages as controversial. Finally, two random baselines which label every article as either controversial or non-controversial based on a coin flip, are presented for comparison (average of three random runs). One of these baselines flips a coin with 50% probability, and the other flips it with 29.8% probability (the incidence of controversy in the training set).

4.5 Results

The results of our approach on the test set are displayed in Table 4.3. For ease of discussion, we will refer to row numbers in the table. For brevity and clarity, highly similar runs are omitted.

4.6 Discussion

As the results in Table 4.3 show, our fully-automated approach (rows 1-11) achieves results higher than all baselines (rows 15-19), in all metrics except recall (which is trivially 100% in row 19).

The parameters that optimized for $F_{0.5}$ on the training set were the best for $F_{0.5}$ as well as Accuracy (row 2), with 10.1% absolute gain in accuracy (16.3% relative gain) over the non-controversial class baseline, which had the best accuracy score among the baselines. For $F_{0.5}$ this run showed 19.5% absolute gain (44.5% relative gain) over the best $F_{0.5}$ score, which was achieved by the Random_{50} baseline.

Even though none of the results displayed in the table were optimized for precision, they still had higher precision than the baselines across the board (compare rows 1-11 to rows 15-19). Among the voting methods, the method that optimized for F_1 on the training set was the Majority voting, using Light Stopping, aggregating over the maximal value of 15 neighbors, with discriminative thresholds for both M and C (row 12). This run showed a 10.4% (18.9% relative gain) absolute gain on the test set over the best baseline for F_1.

The results of the sentiment baseline (row 15) were surprisingly similar to a trivial acceptor of "all controversial" baseline (row 19); at closer look, the sentiment classifier only returns about 10% of the webpages as lacking sentiment, and thus its results are close to the baseline. We tried applying higher confidence thresholds to the sentiment classifier, but this resulted in lower recall without improvement in precision. We note that the sentiment classifier was not trained to detect controversy; it's clear from these results, as others have noted, that sentiment alone is too simplistic to predict controversy [Awadallah et al., 2012b, Mejova et al., 2014].

When comparing our results (rows 1-11) to the best oracle-reliant runs from prior work (rows 12-14, see [Dori-Hacohen and Allan, 2013]), the results are quite comparable. Recall that this supervised work represents a proof-of-concept upper-bound analysis, with a human-in-the-loop providing judgments for the relevant Wikipedia pages, rather than an

automatic system that can be applied to arbitrary pages[7]. When comparing the best supervised result (row 12) to the best weakly-supervised run (row 2) using a zero-one loss function, the results were not statistically different. This demonstrates that our novel, fully-automated system for detecting controversy on the web is as effective as upper-bound, human-mediated predictions.

4.7 Conclusions

In this chapter, we presented the novel problem of detecting controversial topics on the web (web classification of controversy), and contributed a nearest-neighbor approach that presented a first solution to this problem (contribution 1.3.2.1). Our algorithm is based on a K-Nearest-Neighbor classifier that maps from webpages to related Wikipedia articles, thus leveraging the rich metadata available in Wikipedia to the rest of the web.

We first created a system that relied on human judgment for neighbors drawn from Wikipedia and evaluated it (contribution 1.3.2.2). We demonstrated that using a human oracle for determining controversy in Wikipedia articles can achieve an $F_{0.5}$ score of 0.65 for classifying controversy in webpages. We showed absolute gains of 22% in $F_{0.5}$ on our test set over a sentiment-based approach, highlighting that detecting controversy is more complex than simply detecting opinions.

We then presented the first fully automated approach to solving the recently proposed binary classification task of web controversy detection (contribution 1.3.2.3). We demonstrated that our system is statistically indistinguishable from the human-in-the-loop approach it is modeled on, and achieved similar gains over prior work baselines (20% absolute gains in $F_{0.5}$ measure and 10% absolute gains in accuracy). We showed that such detection can be performed by automatic labeling of exemplars in a nearest neighbor classifier.

[7]Note that this is not a strict upper-bound limit in the theoretical sense, but in principle it's reasonable to assume that a human annotator would perform as well as an automated system. In fact, in a few cases the automated system performed better than the oracle-reliant approach, see e.g. F1 on row 8 vs. row 14.

Our approach improves upon our previous work by creating a scalable distantly-supervised classification system, that leverages the rich metadata available in Wikipedia, using it to classify webpages for which such information is not available. We relied on Wikipedia labels using various methods, including the M score (we explore this score further and demonstrate that it can be considered a variation of contention in Chapter 6). That said, our approach is modular and therefore agnostic to the method chosen to score Wikipedia articles; like Das et al. [Das et al., 2013], we can leverage future improvements in this domain. For example, scores based on a network collaboration approach [Sepehri Rad and Barbosa, 2012] could be substituted in place of the **M** and **C** values, or added to them as another feature. Likewise, scores developed in the future that directly evaluated other dimensions of controversy in Wikipedia could be used in addition or instead of the scores used here. The nearest neighbor method we described is also agnostic to the choice of target collection we query; other rich web collections which afford controversy inference, such as Debate.org, Debatabase or procon.org, could also be used to improve precision.

Finally, as a side-effect of our evaluation efforts, we collected and publicly released a data set of 377 web pages and 1761 Wikipedia articles annotated with regards to controversy, which is the first data set available for this new problem, and the largest data set of controversy labels to date (contribution 1.3.2.4). The data set also includes 3430 annotations of pairs of webpages and Wikipedia articles, regarding whether or not the Wikipedia page is on the same topic as the webpage.

CHAPTER 5

COLLECTIVE INFERENCE FOR CONTROVERSY DETECTION IN WIKIPEDIA

As discussed in Chapters 2 and 4, automatically distinguishing between controversial and noncontroversial topics is a challenging problem that would allow many positive and interesting applications, such as alerting users to controversy or visualizing stances on a controversy. Specifically, Wikipedia's rich metadata and edit history offer valuable resources which can be used to automatically detect controversial topics. In particular, as we describe in Chapter 4, these automatically labeled controversy articles in Wikipedia can serve as a valuable source for classifying and detecting controversy on the web. Existing work on controversy detection in Wikipedia focuses on the properties of individual articles taken in isolation [Kittur et al., 2007, Yasseri et al., 2012], or in some cases on the properties of the editors of those articles [Sepehri Rad et al., 2012].

In this problem domain, we have clear evidence that the intensities of controversy among related pages are not independent of each other, and therefore, using the controversy level of related pages may improve inference between Wikipedia and the web (see Chapter 4). This relates to the cluster hypothesis [Rijsbergen, 1979] which proposes that related documents are similar in terms of their information needs. Likewise, homophily is the principle that argues that "similarity breeds connection" [McPherson et al., 2001]; if our hypothesis is true, it would imply homophily among controversial topics in Wikipedia, i.e. that related pages would have similar controversy. In the collective inference literature, as we discussed in Section 2.6, "relations" refer to connections in a relational database, such as hyperlinks (cf. [Abiteboul et al., 2000, pg v]). Here, we extend that definition to notions of relatedness that are influenced from the Information Retrieval community, such

as text similarity. Some recent work has alluded to the possibility that controversies occur in neighborhoods of related topics [Das et al., 2013], including our own work from Chapter 4, but this potential connection has yet to be tested directly, nor used to improve controversy detection within a single domain.

To that end, we first analyze a Wikipedia data set to directly present evidence for homophily. We then employ a collective inference approach, which exploits the dependencies among related pages, and demonstrate that our approach improves classification of controversial web pages when compared to a model that looks at each page in isolation. We therefore demonstrate empirically that controversial topics "run in neighborhoods". Our novel approach for detecting controversy in Wikipedia draws on state-of-the-art research in collective and stacked inference, which outperforms most of the existing methods and performs equivalently to the best approach, despite using language-independent features. Our approach can be further generalized to include additional features from other sources, such that any improvement in the intrinsic classification can be translated to further improvements in stacked classification.

5.1 Homophily with respect to Controversy in Wikipedia

We'd like to find out whether Wikipedia articles do in fact exhibit homophily with respect to controversy, i.e. whether linked articles are more likely than random to share the same controversy label. To that end, we examine the Controversy Annotation Dataset we created, particularly its approximately 2000 labeled Wikipedia pages (see Section 4.2). These articles include hyperlinks to other Wikipedia pages, some of which are also included in our data set. If there is no homophily, we would expect linked pages to be just as likely as randomly chosen pages to have a certain combination of labels. If homophily is present, however, we would expect a very different pattern of label distribution. Homophily would imply that linked pages will be more likely than random to share the same annotation label (either both controversial or both non-controversial).

Let $G_{WP}(V_{WP}, E_{WP})$ be the graph of all pages and links in Wikipedia. Let $A = \{(v, c)|v \in V_{WP} \wedge c \in \{0, 1\}\}$ be a set of binary annotations, where $c = 0$ if v is non-controversial and 1 if it is. Let $V_a \subseteq V_{WP}$ be the subset of annotated pages in our data set, such that $V_a = \{v \in V_{WP}|\exists c, (v, c) \in A\}$. Now, let $E_a \subseteq E_{WP}$ be the subset of edges induced among the pages of our annotated data set, such that $E_a = \{(v_1, v_2) \in E_{WP}|v_1, v_2 \in V_a\}$ is the list of adjacency pairs in the induced subgraph, $G_a = (V_a, E_a)$. Let P_a be the multiset (or bag) of binary annotations for adjacency pairs. In other words, $P_a = \{p = (c_1, c_2)|\exists v_1, v_2 \text{ s.t. } (v_1, c_1), (v_2, c_2) \in A \wedge (v_1, v_2) \in E_a\}$.

To test whether Wikipedia articles exhibit homophily with respect to populations, we perform a randomization test on the graph, which contains 34469 pairs of linked annotations. In order to preserve the graph structure, we cannot permute the annotated adjacency pairs P_a directly, as that would effectively reshuffle the entire graph and change the underlying graph structure.

Instead, we permute the annotation values on the vertices, while preserving the original graph structure. We generate 1000 random permutations $A_1..A_{1000}$, such that for each pair $(v, c) \in A$, v is held constant while c is randomly permuted within the set. Note that V_a and E_a remain unchanged. However, new multisets $P_1..P_{1000}$ are induced, such that $\forall i \in [1, 1000], P_i = \{p = (c_1, c_2)|\exists v_1, v_2 \text{ s.t. } (v_1, c_1), (v_2, c_2) \in A_i \wedge (v_1, v_2) \in E_a\}$.

Now, we can perform a $\tilde{\chi}^2$ test on each set P_i, and compare them to the $\tilde{\chi}^2$ for the original P_a. The definition for $\tilde{\chi}^2$ is as follows: $\tilde{\chi}^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$, where E_k and O_k represent the expected and observed values for type k, respectively. In our case, there are four possible types that k can take, representing the 2x2 possibilities for the binary annotations on the edges: $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$. In order to ensure consistency of the expected and observed values for $(0, 1)$ and $(1, 0)$, throughout the homophily calculation we sort the adjacency pairs by alphabetical order of vertex title.

The null hypothesis is that Wikipedia articles do not exhibit homophily, and thus that the $\tilde{\chi}^2$ of P_a would be similar to the distribution of $\tilde{\chi}^2$ among the permutations. In fact, we

find that the $\tilde{\chi}^2$ for P_α was at the 100th percentile of results among the randomly generated values, which indicates that this distribution is extremely unlikely to have been generated at random. Thus, we can reject the null hypothesis that Wikipedia articles' controversy labels are independent from each other at $p < 0.001$. Rather, it's clear that homophily is in fact exhibited with respect to controversy. This is visually striking when comparing the histograms in Figure 5.1, which depicts the histogram of $\tilde{\chi}^2$ only for the permuted pairs $P_1..P_{1000}$, and Figure 5.2, which is rescaled to include $\tilde{\chi}^2$ for the original P_α.

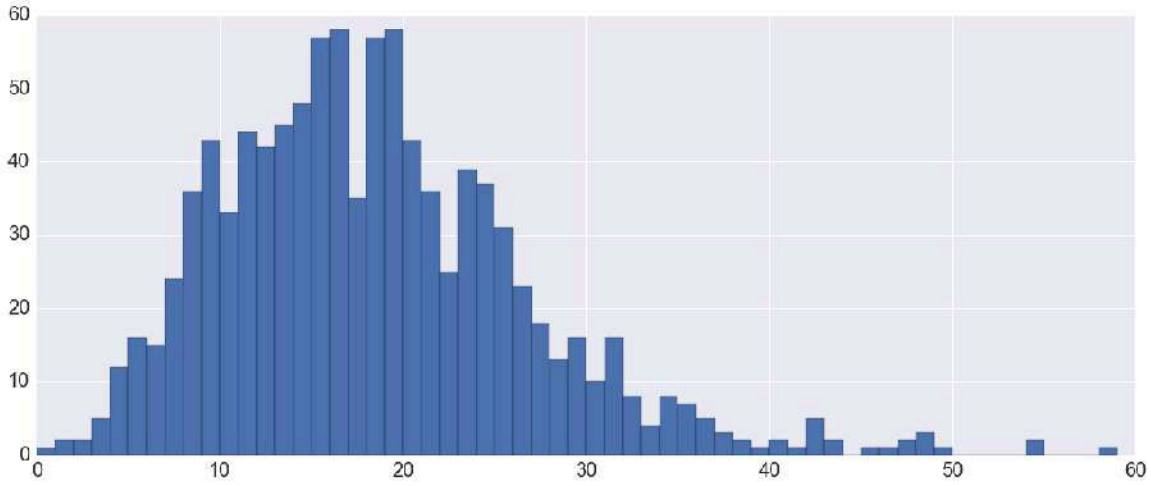

Figure 5.1: Histogram of $\tilde{\chi}^2$ values for $P_1..P_{1000}$.

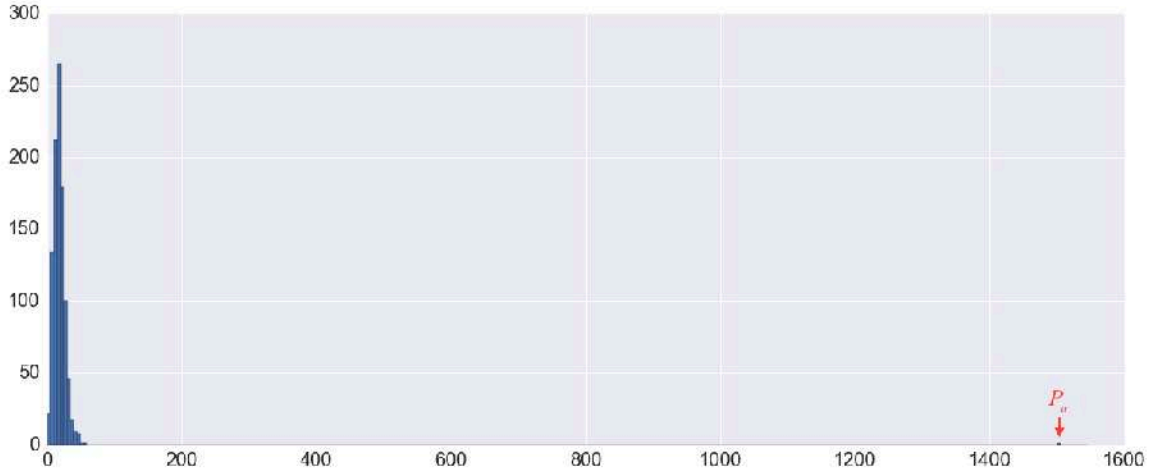

Figure 5.2: Histogram of $\tilde{\chi}^2$ values for $P_1..P_{1000}$ as well as the original P_α (far right).

5.2 Controversy Detection in Wikipedia

We will now classify Wikipedia pages as controversial or not, using a combination of intrinsic features of a page, as well as predictions of controversy from pages related to it (based on the same feature set).

5.2.1 Structure and Intrinsic Features in Wikipedia

As described in Chapters 3 & 4, Wikipedia's collaborative structure allows any user to edit a page, with the entire edit history being recorded. This makes it a unique resource of rich metadata features for the purpose of controversy classification, which are largely unique to Wikipedia; examples include features such as the number of edits performed on a page, the number of registered and anonymous users editing it, and so forth.

Other possible features include a more direct analysis of the actual edits performed in Wikipedia, such as modeling which links were changed on the page [Bykau et al., 2015], or the feedback provided by viewers of that page as a major feature of classifiers [Jankowski-Lorek et al., 2014]. As we will discuss in Section 6.1, it's possible to measure contention by evaluating "edits wars" between users [Yasseri et al., 2012]. However, for ease of comparison with prior work, in this chapter we focus on the features used in the "meta" classifier [Kittur et al., 2007], that were demonstrated in a prior comparative work to achieve the best results in the Wikipedia controversy classification task [Sepehri Rad and Barbosa, 2012]. This classifier is similar to the "C" score described in Chapter 4, rather than to the M score which we will analyze further in Chapter 6.

5.2.2 Diversity of Links in Wikipedia

We would like to examine the neighborhood of each Wikipedia page, for the purpose of using stacked classification and evaluating whether homophily exists for controversial topics in Wikipedia.

The effectiveness of collective inference relies on autocorrelation between related instances. In the terms of our problem, if a page is controversial, then the pages related to it

Algorithm 1 Cross-validation stacked training procedure

for fold $i = 1..k$, $Set_i = A \setminus \text{fold}_i$ **do**
Train IM_i, an intrinsic model on Set_i
Select $subneighbors(Set_i) \subseteq neighbors(Set_i)$
Apply IM_i on $subneighbors(Set_i)$
Aggregate predictions of $subneighbors(Set_i)$ to create an extended feature set, Set'_i
Train SM_i, a stacked collective model on Set'_i

Algorithm 2 Cross-validation stacked inference procedure

for fold $i = 1..k$ **do**
Select $subneighbors(fold_i) \subseteq neighbors(fold_i)$
Apply IM_i (trained above) on $subneighbors(fold_i)$
Aggregate predictions of $subneighbors(fold_i)$ to create an extended feature set, $fold'_i$
Apply SM_i (trained above) on $fold'_i$

are likely controversial, and vice versa. The controversy level of related pages, therefore, can be used as a feature to the collective model.

However, there is a potentially complicating factor - the relational links available in Wikipedia, i.e. hyperlinks, represent a variety of different things; not all represent topical connections, and thus are noisy with regards to topic. In other words, not all links are equally useful for the purpose of stacked classification. For example, the Wikipedia page for the controversial topic "Creationism" has links to pages on related (controversial) topics such as "Creation Science" and "Young Earth Creationism," but it also has links to pages on largely non-controversial topics such as "Newsweek" and "Moon" which don't directly relate to the Creationism topic. We hypothesize that stacked classification to be more useful when applied specifically to those more similar links. We can apply similarity metrics, such as TF-IDF based cosine similarity, to choose which neighbors to use.

In the case of Wikipedia, we also argue that links pointing into, and out of, an article, should be viewed as separate types of relationships, rather than treating them as equivalent. For example, as shown in Figures 5.3 and 5.4, incoming links consist of a Zipfian-like distribution which grows on a logarithmic scale, while outgoing links exhibit a more linear relationship.

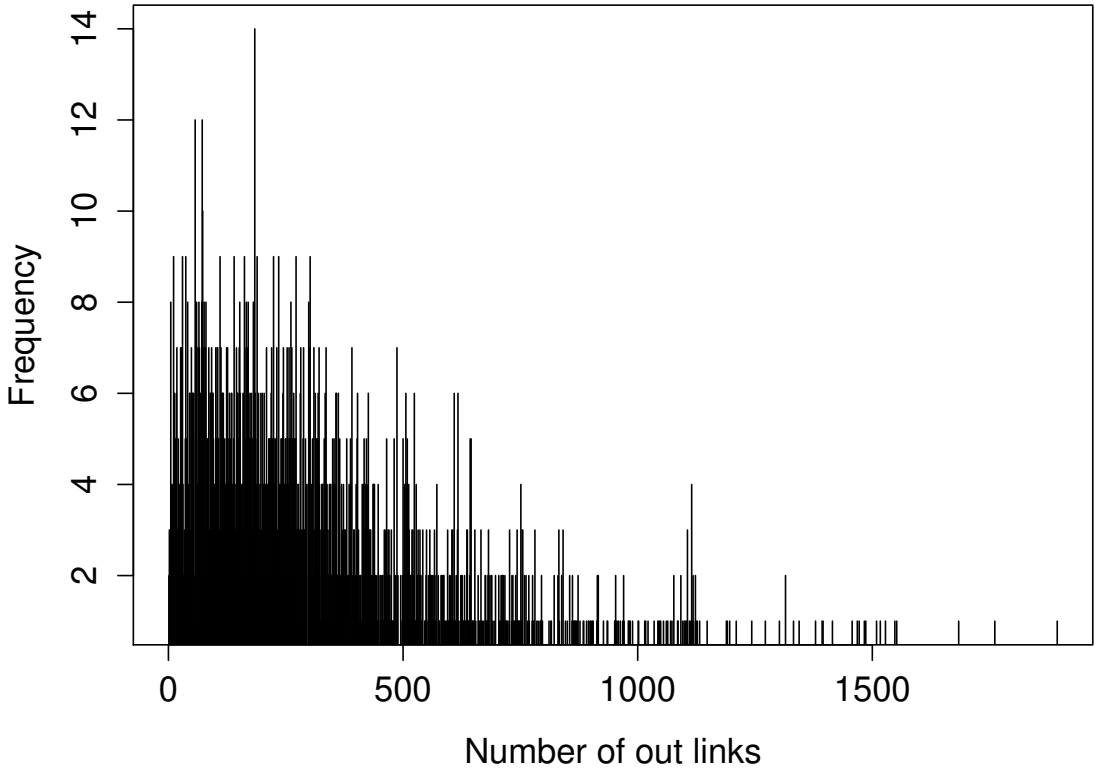

Figure 5.3: Distribution of counts of outgoing links for Wikipedia articles in our data sets at **linear scale**.

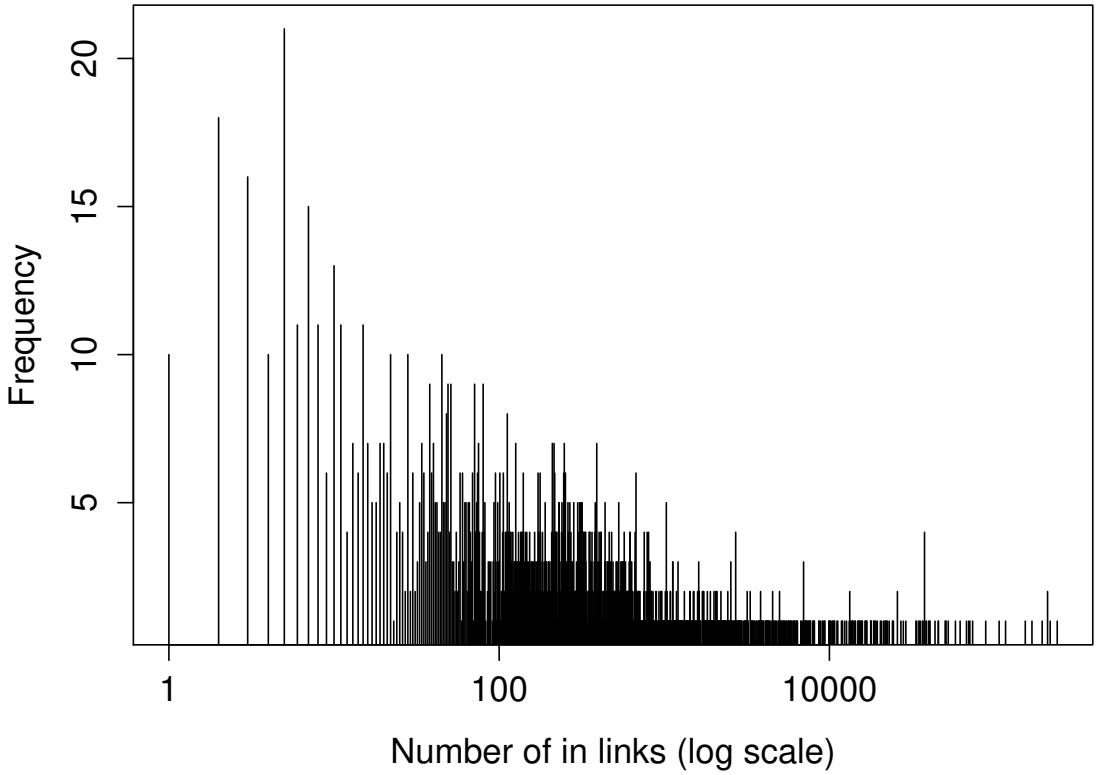

Figure 5.4: Distribution of counts of incoming links for Wikipedia articles in our data sets at **logarithmic** scale.

5.3 Approach

We will classify Wikipedia pages as controversial or not, using a combination of intrinsic features of a page, as well as predictions of controversy from pages related to it. There are two novel parts to our approach (described below): first, we construct a sub-network of relations based on similarity, and then proceed to use a stacked model on top of this constructed network. The training procedure for the intrinsic model is the standard fashion. Following Kou and Cohen [Kou and Cohen, 2007], our stacked training procedure creates neighbor predictions in a cross-validated manner with 10 folds. The main difference from their approach is the use of a subset of the neighbors, rather than all neighbors. The training procedure is applied to the i-th fold, as seen in Algorithm 1. At inference time, the stacked model pipeline is applied to the i-th fold in an analogous manner, as seen in Algorithm 2.

5.3.0.1 Constructing a Sub-network

We examine the neighborhood of each Wikipedia page, for stacked classification and to evaluate whether homophily exists for controversial topics. The effectiveness of collective inference relies on autocorrelation between related instances; presumably, if a page is controversial, then the pages related to it are likely controversial. The controversy level of related pages, therefore, can be used as a feature to the collective model. However, links in Wikipedia are noisy, and not necessarily the best indication of relatedness. We expect stacked classification to be more useful when applied specifically to more relevant links; we thus do not consider every hyperlink to be an equally valid neighbor, but instead apply a similarity function to generate a relative ranking among all neighbors. Specifically, we construct a sub-network by applying a TF-IDF based pairwise cosine similarity function on the text of the page, and then selecting the top-scoring neighbors (taken as two separate lists, for in-links and out-links) as most "related" to the center page.

5.3.0.2 Creating a Stacked Model

To evaluate our hypotheseis, we create intrinsic and collective models of controversy. We compare an *intrinsic classifier* that classifies each page independently, and a **collective inference classifier** that assumes dependence between controversy values of related pages.

5.4 Data set and Experimental Setup

We examine the following hypotheses: (1) Stacked models present an improvement in inference over intrinsic models; (2) using a subset of chosen neighbors, based on a similarity ranking, represents an improvement upon using all neighbors; (3) using this subset also represents an improvement upon using the same amount of random neighbors. We will describe the data sets used, the model features and setup, and finally the alternative systems we created in order to examine our hypotheses.

5.4.1 Data Sets

We use two data sets for this work, as described in Table 5.1, which were created by two independent groups. The first data set is one that we created as part of this thesis, and which we made publicly available[1]: the Controversy Annotation Dataset, which we denote as DHA (see Chapter 4); in particular, we use the annotations of approximately 2000 Wikipedia pages that are included in this set (see Table 5.1)[2]. The Wikipedia articles in the set were created around 40 seed topics, as described in Section 4.2. This data set is not balanced, as it contains about 15% controversial articles.

The second data set is a collection of 480 pages provided by Seperhi Rad et al. ([Sepehri Rad et al., 2012]), which we denote as SRMRB. This data set was selected independently from the DHA data set, by randomly selecting articles that met specific criteria

[1] See ciir.cs.umass.edu/downloads/

[2] The set of Wikipedia articles is slightly larger here due to the usage of about 200 articles that were annotated and released in the Controversy Annotation Dataset, but excluded from the analysis in Chapter 4.

Table 5.1: Data set size and annotations (Wikipedia Articles)

Set	Articles	Controversial
DHA [Dori-Hacohen and Allan, 2013]	1926	293 (15.2%)
SRMRB [Sepehri Rad et al., 2012]	480	240 (50%)

from the set of featured articles in Wikipedia. This data set is exactly balanced, i.e. contains 50% controversial articles and 50% noncontroversial.

While it is quite challenging to estimate the precise incidence of controversy in the wild, we believe that an unbalanced setting, as in the DHA data set, is more realistic - in general, noncontroversial topics far outnumber controversial topics.

In order to partially mitigate the challenges of training on an imbalanced set (DHA), we applied weights to all the instances in the training folds, such that the sum of weights of all controversial pages was equal to the sum of weights for the noncontroversial pages.

5.4.2 Model Features and Setup

For both the intrinsic and the stacked models, we use the Random Forest classifier [Breiman, 2001] provided by Weka, set to use 100 trees. We use the default behavior for all other settings[3]. We chose to use random forest due to its feature selection capabilities. For training and inference, we used 10-fold cross-validation, as described in Section 5.3.

5.4.2.1 Similarity for Sub-network Construction

In order to generate the collective model, we observed all Wikipedia pages linking into, and out of, the center page. We ranked all these pages by pairwise, TF-IDF based cosine similarity (ignoring stop words), then chose the top k in-links and the top k out-links of the central page. We considered several alternatives for thresholding the similarity, but in the experiments described below, we simply pick the top k ranked neighbors for incoming links, as well as the top k for outgoing links, where k is either 10 or 300.

[3]See http://weka.sourceforge.net/doc.dev/weka/classifiers/trees /RandomForest.html

Table 5.2: Intrinsic and Stacked features

Type	Description
Intrinsic	# of Revisions; # of Minor Revisions; # of Editors; # of Anonymous Editors; # of Anonymous Revisions; % of Anonymous Editors; % of Anonymous Revisions; Max Edits Per Editor; Avg Edits Per Editor; Std Dev of Edits Per Editor
Stacked	Proportion of neighbors above X% probability of controversy, where X = {10%, 30%, 50%, 70%, 90%}. This represents a discretized version of the probability distribution of controversy among the neighbors (as scored by the intrinsic classifier); Max Controversy probability among neighbors; Avg Controversy probability among neighbors

5.4.2.2 Features

The features of both of the intrinsic and stacked models are displayed in Table 5.2. *Intrinsic Features* follow prior work that used metadata features of the Wikipedia pages [Kittur et al., 2007, Sepehri Rad and Barbosa, 2012]. All intrinsic features are extracted from the May 2014 Wikipedia dump; [4] a subset of the features were extracted using JWPL[5]. We use the intrinsic model to generate predictions (probabilities of controversy) for each neighbor in the sub-network described above. Collective and stacked inference requires that the relevant features of pages be aggregated in order to use them: we use the aggregate functions listed in the bottom part of Table 5.2, applied separately to in-links and out-links. In total, 14 *Stacked Features* were added (7 aggregates each, which were applied to the top k in-links and out-links separately).

5.4.3 Alternative Systems

Our proposed system described above, which we denote $Stacked\text{-}Ranked\text{-}k$, uses a similarity function to induce a sub-network for the purpose of stacked inference. In order to test our hypotheses, we construct several alternative systems (see Table 5.3); in each case, we train the model on the same intrinsic and stacked features described above (as

[4] https://dumps.wikimedia.org/

[5] https://github.com/dkpro/dkpro-jwpl

Table 5.3: Compared Systems

Name	Description
$Stacked$-$Ranked$-k	Proposed stacked inference system with a similarity-based sub-network
$Intrinsic$	A classifier using only intrinsic features
$Stacked$-All	A stacked inference system, as above, but which uses all Wikipedia neighbors
$Stacked$-$Random$-k	A stacked inference system which uses k randomly selected neighbors
$Neighbors$-$Only$-k	A classifier based only on the neighbor predictions (as in a regular stacked model), without using the intrinsic features of the center page
Prior work	[Brandes et al., 2009, Kittur et al., 2007, Vuong et al., 2008] [Yasseri et al., 2012]; see [Sepehri Rad and Barbosa, 2012] for comparative study

appropriate for that system). Where possible, we compare our results to several baselines from prior work [Brandes et al., 2009, Kittur et al., 2007, Vuong et al., 2008, Yasseri et al., 2012], as reported in a recent comparative study [Sepehri Rad and Barbosa, 2012].

5.5 Results

We discuss some differences in data imbalance between the two data sets and our choice of metrics, and our findings: using similar neighbors improve stacked inference, neighbors can provide good inference even without intrinsic features, and a stacked model outperforms existing classifiers.

5.5.1 Data Imbalance and Metrics

The results of our experiments are displayed in Table 5.4. Due to the unbalanced nature of the DHA data set, neither F1 nor accuracy are representative metrics for classification; thus, we focus most of our subsequent discussion on Area under ROC (AUC), a metric commonly used to evaluate unbalanced sets, as it is insensitive to data set imbalance; we report F1 and accuracy results for comparison with prior work.

5.5.2 Similar Neighbors Improve Results

The predictive power of the stacked model grows with the number of neighbors; results increase substantially within the first 25 neighbors, with diminishing returns afterwards. The Stacked classifier outperforms both the Intrinsic and Neighbor-only models, for both data sets and all metrics presented (see Table 5.4). For most values of k, our proposed system (which chooses neighbors according to a similarity metric), outperforms a random selection of the same number of neighbors, with the difference clearest when a small number of neighbors is used (see Figures 5.5 and 5.6). Not surprisingly, the systems start converging as the number of neighbors approaches all neighbors of the page.

5.5.3 Neighbors Provide Quality Inference Without Intrinsic Features

As expected, each stacked model outperforms its equivalent Neighbors-only version, which ignores the intrinsic features of the page. Interestingly, in some cases the Neighbors-only model outperforms an intrinsic classifier (see Table 5.4), despite not receiving any features of the page itself; further work is needed to examine this phenomenon.

5.5.4 Stacked Models perform comparably to Prior Work

There are some challenges in comparing our results to prior work on controversy detection in the SRMRB data set, chief of which is that our results are reported on a more up-to-date Wikipedia dump (Sepheri Rad and Barbosa [2012] provide a comprehensive comparative analysis of controversy classification). In addition, these results were reported only in terms of accuracy (percent correct) with no AUC or other metrics reported. With these constraints in mind, and as seen in Table 5.5, our results outperform the Basic, bipolarity, and mutual reverts methods - all results as reported in the comparative study [Sepehri Rad and Barbosa, 2012]. Our result of 74.4% is slightly lower than the Meta classifier[6] [Kittur et al., 2007] (75%).

Our result underperforms the reported state-of-the-art Editor Collaboration classifier [Sepehri Rad and Barbosa, 2012, Sepehri Rad et al., 2012] (84%). Unfortunately, we were not able

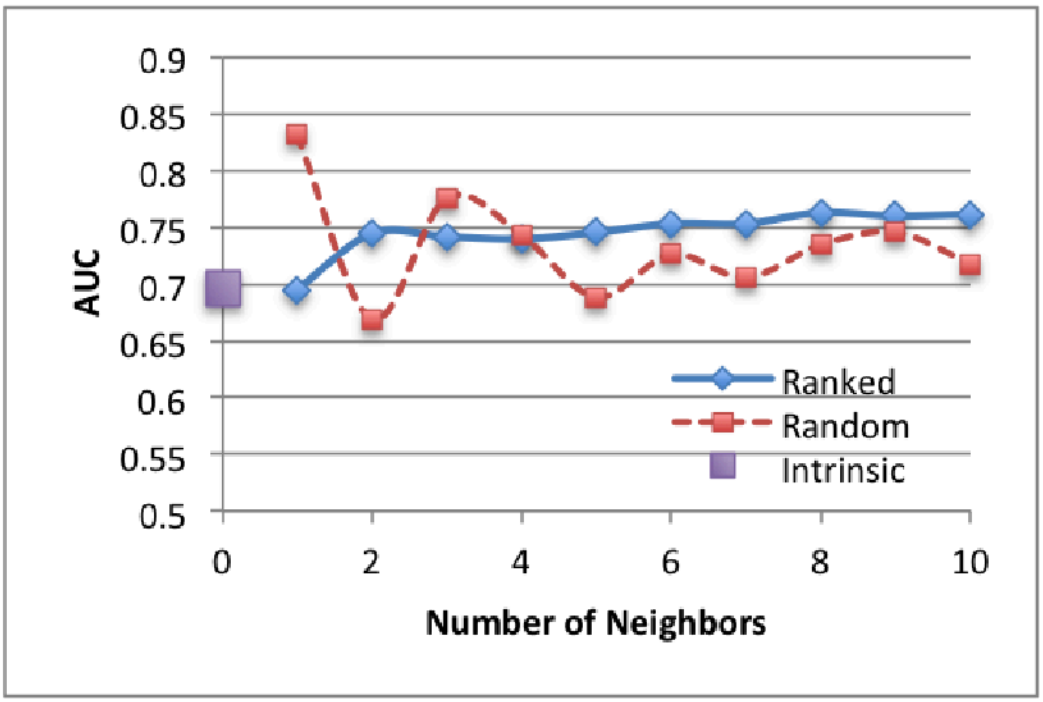

Figure 5.5: AUC as a function of number of neighbors, for those ranked by a similarity metric or selected at random, on the DHA data set

to reproduce the code for the Editor Collaboration classifier, which could have been used as an alternative intrinsic classifier. Notably, stacked models are agnostic to the choice of intrinsic classifier for the problem. Demonstrating the principle of consistent improvement [Oakley and Berlin, 1946],we propose that any future improvement in intrinsic, per-page classification of controversy can be enhanced by applying a stacked classifier on top of it which will consider its surrounding network of related pages. We hypothesize that adding collective classification to the Editor Collaboration classifier would further increase its state-of-the-art results, and leave such exploration to future work.

[6]Our Intrinsic classifier at 69.6% accuracy is the Meta classifier [Kittur et al., 2007] without Talk Page features. While these features may be useful, Talk pages are infrequently used in non-English Wikipedias [Yasseri et al., 2012]; using those features would likely improve the stacked model.

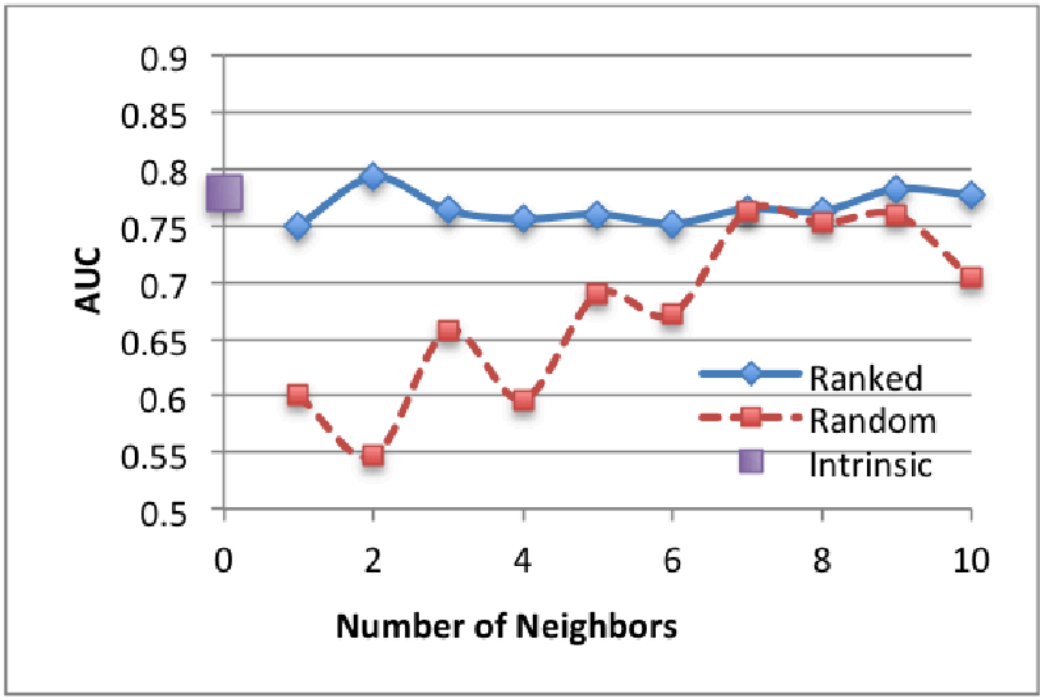

Figure 5.6: AUC as a function of number of neighbors, for those ranked by a similarity metric or selected at random, on the SRMRB data set

Table 5.4: Results for compared models with $k = \{10, 300\}$

Data set	Model	AUC	F1	Acc
DHA	Intrinsic	0.692	0.322	0.788
	NbrOnly-10	0.694	0.244	0.813
	Random-10	0.718	0.289	0.775
	Stacked-10	0.762	0.303	0.823
	NbrOnly-300	0.788	0.348	0.833
	Random-300	0.790	0.367	0.838
	Stacked-300	**0.800**	0.372	**0.844**
	AllNeighbors	0.793	**0.399**	**0.844**
SRMRB	Intrinsic	0.778	0.704	0.696
	NbrOnly-10	0.655	0.620	0.617
	Random-10	0.705	0.697	0.658
	Stacked-10	0.783	0.684	0.670
	NbrOnly-300	0.794	0.704	0.707
	Random-300	0.838	0.736	0.735
	Stacked-300	**0.840**	0.730	0.738
	AllNeighbors	0.828	**0.744**	**0.744**

Table 5.5: Accuracy results compared to prior work. See discussion in text

System	Accuracy
Our Work	74.4%
Basic method [Vuong et al., 2008]	60%
Bipolarity method [Brandes et al., 2009]	56%
Mutual Reverts method [Yasseri et al., 2012]	67%
Meta classifier [Kittur et al., 2007]	75%
Editor Collaboration classifier [Sepehri Rad et al., 2012]	84%

5.6 Conclusion

In this chapter, we address the second theme of this thesis directly: we present evidence for the existence of homophily in Wikipedia with respect to controversy, by demonstrating that the correlation between labels of linked pages was unlikely to be generated randomly. Articles that are linked on Wikipedia are *m*ore likely than random to have the same controversy label, at a 99% confidence interval (contribution 1.3.3.1). We thus show that controversial articles exist in topical neighborhoods of controversy: the cluster hypothesis [Rijsbergen, 1979] and "birds of a feather" principle [McPherson et al., 2001] hold not only to the people editing (as in peer effects, c.f. [Goldsmith-Pinkham and Imbens, 2013]), but also to the level of controversy among the topics linked.

We present a novel algorithm for controversy detection in Wikipedia, based on techniques of collective inference and stacked models, which leverages this homophily to improve inference (contribution 1.3.3.2). We described a stacked model for controversy, which first trains a classifier on the pages' features in isolation, then applies it to predict controversy labels for its neighbors, and finally uses those predictions to estimate the controversy of the center page. Neighbors can be hyperlinks or a subset of hyperlinks that are ranked based on similarity (see below). The resulting stacked model is comparable to prior work results.

Additionally, we presented an advance in the space of relational and collective inference: we demonstrated that a sub-network constructed based on similarity can yield better

classification results than simply taking the default relationship in the relational database (contribution 1.3.3.3). The similar-neighbors model showed improvement over models using all neighbors or randomly selected neighbors. This approach has the potential to improve collective classification results for other problem domains beyond our application to controversy detection, particularly for semistructured data sets such as Wikipedia where text-based similarity metrics can prove valuable for inference. It is an elegant, effective way of incorporating similarity in collective and stacked inference, which is easy to implement in practice and does not require the additional overhead suggested by a more generalized system that can reason about similarity [Bröcheler et al., 2012].

Finally, we also presented a neighbors-only classifier that does not utilize the features of a page itself but only on its neighbors, and demonstrate that, counter-intuitively, in certain cases it can be as effective as a classifier that relies on the page's own features (contribution 1.3.3.4). This interesting result may have implications beyond controversy for the purposes of error-reduction in noisy collections.

CHAPTER 6

CONTENTION ON WIKIPEDIA AND WEB

In order to add one more layer of evaluation for our contention measure (introduced in Chapter 3), this chapter presents a probabilistic contention score on Wikipedia, compare it to a heuristic score from prior work, and evaluate its ability to classify controversy in Wikipedia. We then evaluate the ability to use this contention score for the extrinsic task of detecting controversy on the web. This chapter ties all our work together by bringing our theoretical model (Chapter 3) to bear on the challenging controversy classification tasks in Wikipedia (Chapter 5) and the web (Chapter 4).

6.1 Contention in Wikipedia: re-deriving the M measure

In order to provide further validation for the contention model introduced in Chapter 3, we now turn to evaluating it on Wikipedia, where we have ground truth data for controversy on approximately 2400 topics. Prior work [Sumi et al., 2011, Yasseri et al., 2012] established reverts as a central mechanism for disagreement and controversy in Wikipedia, and introduced a revert-based heuristic for controversy called M (which we used in Chapter 4). Motivated in part by the success of that measure, we develop an alternative computational model of contention that leverages the Wikipedia edit history and a special class of edits called reverts to derive a probabilistic estimate of contention.

6.1.1 Preliminary definitions

In Wikipedia, anyone can edit an article, and the entire revision history is recorded. Since anyone can edit, special types of edits called "reverts" are sometimes used to return

an article to a former version, in case recent changes are unacceptable to some editors. Some preliminary definitions will help us formalize the connection between contention and Wikipedia edits, including reverts, as well as assist in comparing to prior work [Sumi et al., 2011].

Let $\mathfrak{W} = \{D\}$ be the collection of articles in Wikipedia. Let an edit be defined as a pair, $e = (\delta, p)$, such that δ is a set of changes to a document (such as insertions, deletions, and substitutions) and p is the person (editor) that instituted this change (a similar formalism was introduced by Maniu et al. [2011]).

For $D \in \mathfrak{W}$ let $V_D = \{v_0, v_1, v_2, ...v_k\}$ be the set of $k+1$ revisions (or versions) that D goes through, where v_0 is the empty document and $D = v_k$ at present (or at the time in which the Wikipedia snapshot was taken). Let $E_D = \{e_1, e_2, ...e_k\}$ be the set of k edits applied to the document D, with $e_i = (\delta_i, p_i)$, such that applying δ_i to v_{i-1} yields v_i. Note that neither the v_i's nor the e_i's need to be distinct.

Let $\omega_D = \{p \in \Omega | \exists \delta, (p, \delta) \in E_D\}$ be the set of people who created the edits in E_D (also called editors). Likewise, let

$$\Omega_{\mathfrak{W}} = \bigcup_{D \in \mathfrak{W}} \omega_D$$

be the set of all editors in Wikipedia.

Let $R_D \subseteq E_D$ be the set of **reverts**, defined as $R_D = \{e_j \in E_D | \exists i < j-1 \text{ s.t. } v_i = v_j\}$. For simplicity, we ignore any no-op edits in which consecutive versions are identical. In other words, p_j (the author of edit j) made an edit that set v_j to be equal to a prior version, v_i, with at least one intermediate edit discarded completely. In that case, we consider e_j to be a **revert** between the two editors p_j and p_{i+1}, since p_j discarded the edits after p_i's version. (Note that all edits e_k, $i+1 < k < j$ were also discarded. For simplicity, and following Sumi et al. [2011], we only refer to p_{i+1} as being reverted rather than all p_k's.) We denote this as a directed relationship: $reverted(p_j, p_{i+1})$.

6.1.2 The contention definition applied to Wikipedia

The original definition relied on stances and a binary *conflicts* function between any two stances, *conflicts*: $S \times S \mapsto \{0, 1\}$, and can be applied for Wikipedia documents as follows:

$$P(c|\Omega, D) = P(p_1, p_2 \text{ selected randomly from } \Omega, \exists s_i, s_j \in S, \\ \text{s.t. } holds(p_1, s_i, D) \wedge holds(p_2, s_j, D) \wedge conflicts(s_i, s_j))$$

In leiu of that definition, which requires stances, we change the *conflicts* function to be defined between two people directly, such that $conflicts : \Omega \times \Omega \mapsto \{0, 1\}$. Then, the contention definition becomes:

$$P(c|\Omega, D) = P(p_1, p_2 \text{ selected randomly from } \Omega \wedge conflicts(p_1, p_2))$$

In other words, we make a subtle change to the contention model such that it focuses on conflicts between people, rather than on the stances that they hold.

6.1.3 Conflicts and Reverts

Rather than estimating stances, our challenge now becomes to provide an estimate for the *conflicts* function directly between pairs of people. Several past researchers have noted the centrality of Wikipedia reverts to the study of controversies [Brandes et al., 2009, Sumi et al., 2011, Yasseri et al., 2012]. Yasseri et al. in particular established reverts as a central mechanism for detecting controversy-related disagreement in Wikipedia [Yasseri et al., 2012].

One approach might be to simply consider any revert to represent a *conflicts* relationship. Let $conflicts_r(p_1, p_2) \equiv reverts(p_1, p_2) \vee reverts(p_2, p_1)$, in which case we get:

$$P(c|\Omega, D) = P(p_1, p_2 \text{ selected randomly from } \Omega \wedge (reverts(p_1, p_2) \vee reverts(p_2, p_1)))$$

Unfortunately, this simple approach is likely to be too naïve. We can conceptually distinguish between two types of reverts: those reverting vandalism and those reflecting opposing stances. Vandalism is thus a confounding factor for controversy, which led to problematic results in some past work that was not able to distinguish between the two [Vuong et al., 2008].

A reasonable implementation choice is to use non-vandalism reverts as an estimation of the *conflicts* relationship. Sumi, Yasseri and their colleagues argued that non-vandalism reverts are prevalent for controversial topics, and claimed that vandalism reverts were fairly easy to distinguish from non-vandalism (i.e. true controversy) reverts using a few heuristic approaches [Sumi et al., 2011, Yasseri et al., 2012]. The first heuristic they proposed was to focus exclusively on **mutual reverts**, i.e. cases in which both editors have reverted each other. Let $conflicts_{mr}(p_1, p_2) \equiv reverts(p_1, p_2) \wedge reverts(p_2, p_1)$. Incorporating this heuristic into the definition of conflicts, we get a slightly different formulation:

$$P(c|\Omega, D) = P(p_1, p_2 \text{ selected randomly from } \Omega \wedge conflicts_{mr}(p_1, p_2)) =$$
$$P(p_1, p_2 \text{ selected randomly from } \Omega \wedge (reverts(p_1, p_2) \wedge reverts(p_2, p_1)))$$

However (again according to Sumi et al. [2011]), even mutual reverts are not sufficient to eliminate vandalism reverts completely. They devised a reputation factor per editor, which grows proportionally with the number of edits the user contributes to this specific article. The likelihood of an editor being a vandal is independent of all other editors. Adopting a probabilistic approach, we can re-formulate the *conflicts* relationship, rather than being a binary value, into a probabilistic expression that captures the likelihood of a pair of editors reverting each other without vandalism. We can express this probability conditional on the existence of a mutual revert, as such:

$$P(conflicts(p_1, p_2)|reverts(p_1, p_2) \wedge reverts(p_2, p_1)) = P(p_1 \text{ is not a vandal}) \\ * P(p_2 \text{ is not a vandal})$$

and:

$$P(conflicts(p_1, p_2)|\neg reverts(p_1, p_2) \vee \neg reverts(p_2, p_1)) = 0$$

Note that relying solely on non-vandalism mutual reverts is likely to be an underestimate of the *conflicts* relationship, since editors can (and in fact, likely do) have conflicting stances without ever reverting each other. In practice, however, this underestimate holds across Wikipedia, and for the purposes of the analysis, we can assume that the rank order of contention in these topics is unlikely to be seriously impacted by this underestimate.

In order to progress further, we need to estimate the probability that a specific person p is (or is not) a vandal. Here, indirectly following Sumi et al.'s reputation factor, we choose to use the number of edits a user has contributed to E_D, divided by the largest reputation factor for any editor on the page. To restate this formally, let

$$E_{p,D} = \{e \in E_D | \exists \delta, e = (p, \delta) \in E_D\}$$

be the set of edits contributed to document D by editor p. Let $N_p^D = |E_{p,D}|$ be the size of said set, i.e. the number of edits contributed to D by p. Let

$$N_{max}^D = \max_{p \in \omega_D} N_p^D$$

Now, we estimate the probability of p's non-vandalism as:

$$P(p \text{ is not a vandal}) = \frac{N_p^D}{N_{max}^D + 1}$$

Note that this probability is independent for each editor, and is in the range $[\frac{1}{N_{max}^D+1}, \frac{N_{max}^D}{N_{max}^D+1}]$.

We can marginalize over all pairs of editors for the document, and incorporate this probability into our contention estimate. Let $MR_D = \{(p_i, p_j) | p_i, p_j \in \omega_D \text{ s.t. } i < j \wedge reverts(p_1, p_2) \wedge reverts(p_2, p_1)\}$ be the set of pairs that have mutual reverted each other. Then we can calculate contention as follows:

$$P(c|\Omega, D) = \frac{\sum_{p_1, p_2 \in \omega_D} P(conflicts(p_1, p_2))}{|\Omega|^2} =$$

$$\frac{1}{|\Omega|^2} * \sum_{(p_i, p_j) \in MR_D} P(p_i \text{ is not a vandal}) * P(p_j \text{ is not a vandal}) =$$

$$\frac{1}{|\Omega|^2} * \sum_{(p_i, p_j) \in MR_D} \frac{N_{p_i,D}}{N^D_{max} + 1} * \frac{N_{p_j,D}}{N^D_{max} + 1}$$

Note that we select the editors from ω_D, yet we can measure contention over any superset of ω_D, for example $\Omega_{\mathfrak{W}}$. This allows us to compare contention across either local (article-specific) populations as well as larger ones, up to and including all of Wikipedia's editors.

6.1.4 The original definition of M

We can now compare our contention-based model to the original definition of M [Sumi et al., 2011]. As discussed above, Sumi et al. considered two main heuristic factors for differentiating vandalism from non-vandalism reverts. First, they considered pairs of editors that were mutually reverting, i.e., that each editor had reverted the other. Second, they defined $N_{p,D}$ to be a reputation factor for an editor, and heuristically used $min(N_{p_i,D}, N_{p_j,D})$ as the reputation for the pair of editors. Under our formalism, and despite the likely independence of the editors' chances of being vandals, they manually set $P(p_1 \text{ is not a vandal} \wedge p_2 \text{ is not a vandal}) = min(N_{p_i,D}, N_{p_j,D})$.

Then, they defined an intermediate measure, which they call M_r, as follows [Sumi et al., 2011]:

$$M_r = \sum_{(p_i, p_j) \in MR_D} min(N_{p_i,D}, N_{p_j,D})$$

Table 6.1: Statistics on the English Wikipedia data set

	# Articles	Total Edits	Revert Edits	Non-revert Edits
English Wikipedia	4,644,479	219,851,361	22,277,895	197,573,466

Their final formulation for M had an additional heuristic embedded in its formula. Let $\omega_D^R = \{p_i \in \omega_D | \exists p_j, (p_i, p_j) \in MR_D \vee (p_j, p_i) \in MR_D\}$ denote the set of all editors that have ever *mutually reverted* on the page. Then Sumi et al. defined their the original M value as:

$$M = |\omega_D^R| \times M_r = |\omega_D^R| \times \sum_{(p_i,p_j) \in MR_D} min(N_{p_i,D}, N_{p_j,D})$$

We can see this score is monotonically increasing for a page over time, and technically unbounded. This is in contrast to our probabilistic score which is can go up or down over time (e.g. if more editors join and do not add mutual reverts) and is bound between [0,1]. What is clear is that both these scores are effectively attempts to quantify contention. We used the M score previously in Chapter 4 as a measure of contention for Wikipedia, where it provided a source of labels for a distantly supervised model of controversy on the web.

6.2 Evaluation

In order to evaluate our scoring approach and compare it to the M score, we generated the M scores as well as two contention scores for each article in the English Wikipedia[1]. Some statistics on the data set are provided in Table 6.1. Approximately 10% of all edits in the data set are reverts.

For contention, in contrast with M, any number of contention scores could in principle be formulated for any number of different populations of Wikipedia editors. A few examples are "editors who have edited articles in the science category," "editors who have followed articles for pop singers," "editors who have edited this document and at least two others in the same category," "all Wikipedia editors with verified accounts," or "all Wikipedia administrators". For simplicity in this analysis, we focus only on two popula-

tions, ω_D and $\Omega_{\mathfrak{W}}$, and their respective contention scores: $P(c|\omega_D, D)$ and $P(c|\Omega_{\mathfrak{W}}, D)$, i.e., contention measured on the population of the page's editors and on the population of all Wikipedia editors, respectively. Note that ω_D changes for each article. For brevity, we refer to $P(c|\omega_D, D)$ as C_D and to $P(c|\Omega_{\mathfrak{W}}, D)$ as $C_{\mathfrak{W}}$ (which should not be confused with the "C" score described in Chapter 4).

Intuitively, we expect $C_{\mathfrak{W}}$ and M for any given page to be more similar to each other than either of them would be to C_D, since C_D is affected more strongly by the number of editors per page which changes in every page.

We evaluate these three scores (M, C_D and $C_{\mathfrak{W}}$) for two tasks for which we have ground truth: the controversy classification task in Wikipedia (see Chapter 5), and the controversy classification task in the web (see Chapter 4).

In order to further explore the difference between the scores, we proceed with a qualitative analysis of the top-scoring articles in each score, including many that were not included in our data sets.

6.2.1 Evaluation on classification tasks in Wikipedia and Web

As mentioned above, we evaluate these three scores (M, C_D and $C_{\mathfrak{W}}$) for two tasks for which we have ground truth:

- Controversy classification task in Wikipedia, for which we have a data set of about 2400 labeled Wikipedia articles (from combining both data sets presented in Chapter 5; see Table 5.1).

- Controversy classification task in the web, for which we have a data set of about 300 labeled web pages (as presented in Chapter 4; see section 4.2 and 4.1). For this task, we variously use the M, C_D and $C_{\mathfrak{W}}$ controversy scores on related Wiki-

[1]For the purposes of this analysis, we use the light dump data provided from Yasseri and colleagues: `http://wwm.phy.bme.hu/light.html`.

Table 6.2: Results for the Wikipedia classification task on a set of labeled Wikipedia articles, using three different contention scores

Wikipedia Score	AUC
M	0.630
C_D	0.628
$C_{\mathfrak{W}}$	0.624

Table 6.3: Results for the web classification task with three different contention scores for Wikipedia pages

Wikipedia Score	F_1	Acc
M	0.54	0.74
C_D	0.57	0.59
$C_{\mathfrak{W}}$	0.57	0.59
All non-controversial	0	0.62
All Controversial	0.55	0.38

pedia articles, as distantly-supervised labels for the web articles, using a k-Nearest-Neighbor classifier.

First, we examine the performance of the three scores on the task of classifying whether Wikipedia articles are discussing controversial topics. For this task, we measured the area under the ROC curve (AUC) for each score. As seen in Table 6.2, all three scores achieve similar AUC results in classifying controversy on our labeled data set. This suggests that there is little difference in discriminative ability between the three scores on this data set.

We next examine the use of the three scores for the extrinsic task of web classification in controversy, i.e. the task explored in Chapter 4. For this evaluation, we use the data set and use a similar experimental setup (for full description, refer to Chapter 4). First, we find the k nearest neighbors of the web page in Wikipedia using a query with the top ten frequent terms (excluding stop words). In each experiment, we apply one of the three scores to the Wikipedia articles. Then, we aggregate the controversy scores of those k articles to get the final scores for the web page. Parameters are trained in a cross-fold validation manner. For

ease of comparison, we do not use voting between scores. Instead, we use a single score for each Wikipedia page, with three experimental conditions depending on the score used (M, C_D or $C_{\mathfrak{W}}$).

As demonstrated in Table 6.3, there was a slight increase in F_1, but a large reduction in Accuracy when using the probabilistic scores. We hypothesize that this is impacted by the unbalanced nature of this data set, though more exploration would be needed in order to test this hypothesis.

In order to better understand the differences between the scores, we turn to a qualitative analysis of the overall results for these three scores.

6.2.2 Qualitative analysis

As an additional evaluation step, we perform a qualitative analysis of the overall results, including articles for which we do not have a clear ground truth. Let X^Y denote the set of articles that ranked top Y by score X. Then M^{1000}, $C_{\mathfrak{W}}^{1000}$, and C_D^{1000} represent the top 1000 ranked articles in all Wikipedia based on M, $C_{\mathfrak{W}}$ and C_D, respectively.

First, we calculate the overlap in the top 1000 ranked articles of all three scores. Figure 6.1 shows the overlap between these three sets. Notably, $|M^{1000} \cap C_{\mathfrak{W}}^{1000}| = 526$, or in other words, a little over half of the top 1000 ranked articles in the M and $C_{\mathfrak{W}}$ scores are shared between the two sets. This aligns well with our intuition above. We also analyze the rank differences among those topics in $M^{1000} \cap C_{\mathfrak{W}}^{1000}$. The average rank difference was 252.2. Of the 526 articles in the set, 55.5% (n=292) of the articles ranked higher by $C_{\mathfrak{W}}$ than M, with an average of 262.4 in rank; 44.1% (n=232) of them ranked higher by M with an average of 241.5 in rank; and 0.4% (n=2) ranked equally on both scores.

Additionally, as we expected, C_D^{1000} has little to no overlap with the top 1000 articles in the other two scores, with only 3 articles total overlapping between $C_{\mathfrak{W}}^{1000}$ and C_D^{1000}. When analyzing these sets, we also found a difference in the average number of edits and reverts per article and the ratio between them, depending on which score was used (see

Figure 6.1: Venn diagram of overlap between the top 1000 ranked articles according to each score. $|M^{1000} \cap C_{\mathfrak{W}}^{1000}| = 526$; $|C_{\mathfrak{W}}^{1000} \cap C_D^{1000}| = 3$; $|M^{1000} \cap C_D^{1000}| = 0$.

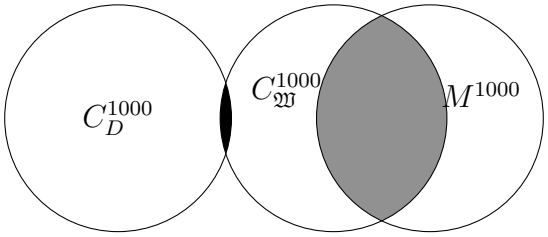

Table 6.4: Edit Statistics on the M^{1000}, $C_{\mathfrak{W}}^{1000}$ and C_D^{1000} sets and their combinations

Set	Avg. Edits	Avg. Reverts	Reverts-to-edits ratio
M^{1000}	6782.2	1413.4	0.21
$C_{\mathfrak{W}}^{1000}$	4741.5	1022.5	0.22
C_D^{1000}	54.9	8.6	0.16
$M^{1000} \cap C_{\mathfrak{W}}^{1000}$	6831.8	1459.7	0.21
$C_{\mathfrak{W}}^{1000} \cap C_D^{1000}$	226.0	40.0	0.18
$C_{\mathfrak{W}}^{1000} \setminus M^{1000}$	2426.2	538.2	0.22
$M^{1000} \setminus C_{\mathfrak{W}}^{1000}$	6727.28	1362.05	0.20

Table 6.4). Notably, the average number of edits and reverts is much higher for M than the other two scores, which makes sense given its monotonically increasing score: articles can only go up in score over time as the number of edits increase, never down, whereas the contention scores are normalized by the number of editors in the population, and could thus go down as the number of editors in the population increase. Additionally, articles in $C_{\mathfrak{W}}^{1000}$ have a slightly higher reverts-to-edits ratio than those in M^{1000}, and both of them have a much higher ratio than those in C_D^{1000}. It's possible that some smoothing is needed in the normalization factor to the C_D score in order to avoid an over-representation of articles with an extremely small population and a disproportionate amount of reverts.

Next, we turn to examine the actual articles that ranked in the top 1000 for M and $C_{\mathfrak{W}}$ in a variety of ways. First, we will examine articles that ranked highly on one list but were not included in the top 1000 in the other. Second, we will examine articles that made the

top 1000 on both lists, but ranked highly on one list and much lower on the other. Finally, we will examine a sample of topics that ranked highly on both scores.

Tables 6.5 and 6.6 show the articles that ranked in the top 100 articles according to one score, but were not included in the top 1000 in the other. Three obvious controversial articles include "Catholic Church" (which includes several subsections on controversial topics such as contraception, homosexuality, etc.), "People for the Ethical Treatment of Animals" (a controversial animal rights group), and "Murders of Channon Christian and Christopher Newsom" (a brutal rape and murder case that was controversial due to white supremacy groups claiming it was a hate crime). A number of these articles relate to pop culture items (celebrities, TV shows or games) that may or may not in fact be controversial; for example, "Michael Jackson" is a clearly controversial figure, whereas "Doctor Who" is clearly not. In a different situation, "Mariah Carey" is a celebrity who has occasionally been controversial; however, a cursory review of the Wikipedia article suggests that many disputes on the article relate to her disputed year of birth, which is variously reported as 1969 or 1970. Interestingly, "Spinosaurus" is an article about a genus of dinosaur that has historically generated a variety of scientific disputes. Overall, we see that each of the scores has a few successes and a few failures, with M leaning towards higher-profile and highly-edited pages.

We can also examine articles that made the top 1000 on both lists, but with a large difference in rank between the two lists. Tables 6.7 and 6.8 show items that ranked in the top 50 in one list and having at least 100 difference in ranks between the two lists. Here, we also see several clear controversial topics on both sides: for example, "List of scientists opposing the mainstream scientific assessment of global warming", "Chronic fatigue syndrome" (a chronic illness surrounded by diagnostic and medical controversies) and "L. Ron Hubbard" (the founder of scientology), rank higher according to $C_{\mathfrak{M}}$, whereas "Barack Obama", "United States" and "Muhammad" rank higher according to M. Likewise, there are prominent articles in both lists that, from a cursory review of the Wikipedia articles, are

Table 6.5: Topics that were in top 100 rank by M but not in the top 1000 by $C_{\mathfrak{W}}$

Article	Rank by M
Michael Jackson	16
Deaths in 2007	20
Catholic Church	36
List of Omnitrix aliens	44
Deaths in 2009	51
Doctor Who	59
Britney Spears	75
Mariah Carey	79
People for the Ethical Treatment of Animals	94

Table 6.6: Topics that were in top 100 rank by $C_{\mathfrak{W}}$ but not in the top 1000 by M

Article	Rank by $C_{\mathfrak{W}}$
Half-Life 2	57
Spinosaurus	59
Murders of Channon Christian and Christopher Newsom	71
SummerSlam (2008)	74
Biff Rose	84
Road to Germany	91

more likely the targets of vandalism than actual controversy, such as "Wii" (rank 10 by M) and "WrestleMania XXIV" (rank 16 by $C_{\mathfrak{W}}$), among others.

As before, there are some obvious hits and misses in each list. As of this writing, many (though by no means all) of the articles in these lists are semi-protected, meaning only registered Wikipedia users can edit them - a protection that is often applied to biographies of living persons who have had recent media exposure[2], as well as manually applied to heavily vandalized or controversial pages. Others yet are noted as under arbitration or manually tagged as controversial. Overall, and along with the previous results, this points

Table 6.7: Topics that were in top 1000 rank on both lists, while scoring in top 50 of $C_{\mathfrak{W}}$ and more than a 100 difference in rank between the lists

Article	Rank by $C_{\mathfrak{W}}$	Rank by M
List of scientists opposing the mainstream scientific assessment of global warming	5	235
Chronic fatigue syndrome	6	153
L. Ron Hubbard	7	194
Moldovans	9	574
List of living supercentenarians	11	340
WrestleMania XXIV	16	125
Taylor Swift discography	18	282
Antisemitism	19	396
Rukia Kuchiki	25	674
John F. Kennedy International Airport	30	217
Tyrannosaurus	32	450
Horcrux	33	466
Newcastle United F.C.	34	216
List of topics characterized as pseudoscience	36	378
List of Avatar: The Last Airbender episodes	37	224
Islamic terrorism	38	505
Ivy League	39	774
Billy Ray Cyrus	40	508
Ireland	46	258
Green Day	49	210

out the weakness in both the M and the $C_{\mathfrak{W}}$ scores in confusing vandalism with controversy due to their heavy reliance on reverts.

Finally, Table 6.9 shows the topics that ranked in the top 100 by both measures. With a few notable exceptions (e.g. "Blackout (Britney Spears album)" and "Naruto Uzumaki", an anime/magna character), most of the articles who made this list are clearly highly controversial. We can hypothesize that the slight difference in the M and $C_{\mathfrak{W}}$ scores' approaches

[2]See `http://en.wikipedia.org/wiki/Wikipedia:Protection_policy#Semi-protection`.

Table 6.8: Topics that were in top 1000 rank on both lists, while scoring in top 50 of M and more than a 100 difference in rank between the lists

Article	Rank by $C_{\mathfrak{W}}$	Rank by M
Barack Obama	589	5
United States	411	6
Muhammad	725	7
Wii	159	10
Intelligent design	246	11
Anarchism	957	14
Akatsuki (Naruto)	182	15
Circumcision	129	17
Jehovah's Witnesses	265	18
John Cena	281	22
Deaths in 2008	428	24
2006 Lebanon War	208	25
Israel	183	27
The Beatles	489	33
European Union	224	34
List of Total Nonstop Action Wrestling employees	273	41
Canada	197	45
Scientology	571	46

to estimating vandalism are capturing different groups of vandals, and by scoring highly on both scores we are able to eliminate more vandalism than each score individually.

6.3 Conclusions

In this chapter, we introduced a variation of our theoretical model of contention by using it to derive a probabilistic, population-based contention score (contribution 1.3.4.1) that is similar in nature to an existing heuristic score, the M measure [Sumi et al., 2011]. We evaluate our new scores on a data set for the Wikipedia task using about 2400 Wikipedia topics, as well as on the web task with about 300 web pages (contribution 1.3.4.2), and find the results are comparable to prior work. We also evaluate our scores in a few

ways, including qualitatively, in comparison to M, and we find a qualitative difference in the scores generated based on which population is evaluated: the editors of a certain page, or all editors on Wikipedia. We find M to be somewhat tilted towards highly-edited (and presumably popular) articles. Our contention score offers two advantages over M: a deeper understanding of what controversy means, as well as its ability to adjust the contention level based on the population being observed.

Table 6.9: Topics that ranked in the top 100 by both measures

Article	Rank by C	Rank by M
George W. Bush	1	2
Super Smash Bros. Brawl	2	8
Avatar: The Last Airbender	3	82
Chiropractic	4	19
List of World Wrestling Entertainment employees	8	1
International recognition of Kosovo	10	72
Transnistria	12	65
Islam	13	12
Global warming	14	3
2009	15	53
Naruto Uzumaki	17	35
Adolf Hitler	20	30
Scotland	21	58
2008	22	29
Eurovision Song Contest 2009	24	87
Falun Gong	26	32
Jesus	27	4
Islamophobia	28	73
Ann Coulter	29	56
Lost (TV series)	35	48
Ayn Rand	41	42
India	42	9
Homeopathy	43	21
Hamas	44	95
September 11 attacks	53	31
Blink-182	54	88
Christianity	58	13
List of social networking websites	61	38
Kosovo	65	40
Blackout (Britney Spears album)	67	61
Sasuke Uchiha	68	85
Organization XIII	79	26
Prem Rawat	81	39
2008 South Ossetia war	85	84
Joseph Smith, Jr.	88	37
Race and intelligence	89	28
Israel and the apartheid analogy	93	66
The Church of Jesus Christ of Latter-day Saints	95	93
The Dark Knight (film)	97	43
RuneScape	99	23

CHAPTER 7

DISCUSSION, CONCLUSIONS & FUTURE WORK

In this thesis, we significantly advanced the computational definition of controversy and the problem of controversy detection, both in Wikipedia and on the web. There were two overarching themes in this thesis:

1. Controversy (or a dimension of it) can be modeled effectively as disagreement within populations.

2. Controversy runs in topical neighborhoods, i.e. it exhibits homophily.

Specifically, in Chapter 3, we defined a novel measure we call "contention", which captures disagreement and is defined with respect to a topic and a population. Our framework defines contention in terms of its topic, but also in terms of the population being observed. We modeled contention from a mathematical standpoint and validated our model by examining a diverse set of sources: real-world polling data sets, actual voter data, and Twitter coverage on several topics. We demonstrated that the contention measure holds explanatory power for a wide variety of observed phenomena that cannot be explained under previous global controversy views. Among these observed phenomena are topics that are well within scientific consensus yet disputed in the general public; polling variations in controversy among certain populations or interest groups; and topics that are controversial only in certain geographical regions or among certain interest groups. We also defined a variation of the contention model which can be applied to Wikipedia, based on special types of edits called reverts, and demonstrated that the M score from prior work [Yasseri et al., 2014] is an approximation of contention. Finally, we empirically demonstrated that contention

is a dimension of controversy, and presented preliminary evidence suggesting that it exists alongside other dimensions, such as "importance". We extended our population-dependent model to this multi-faceted definition of controversy.

In Chapter 4, we posed the novel problem of detecting controversial topics on the web, and constructed the first algorithm addressing it. Our algorithm is based on a K-Nearest-Neighbor classifier that maps from webpages to related Wikipedia articles, thus leveraging the rich metadata available in Wikipedia to the rest of the web. We demonstrated that using a human oracle for determining controversy in Wikipedia articles can achieve an $F_{0.5}$ score of 0.65 for classifying controversy in webpages. We showed absolute gains of 22% in $F_{0.5}$ on our test set over a sentiment-based approach, highlighting that detecting controversy is more complex than simply detecting opinions. We also constructed a fully automated system for web classification of controversy that relies on automated scoring of Wikipedia articles. We demonstrated that our system is statistically indistinguishable from the human-in-the-loop approach it is modeled on, and achieves similar gains over prior work baselines (20% absolute gains in $F_{0.5}$ measure and 10% absolute gains in accuracy).

Chapter 4 implicitly demonstrated our second theme that controversy runs in topical neighborhoods. We made that theme explicit in Chapter 5, where we directly demonstrate that Wikipedia articles exhibit homophily with respect to controversy. In other words, pages that are linked on Wikipedia are more likely than random to have the same controversy label at 99% confidence interval. We then presented a novel algorithm for controversy detection in Wikipedia, based on techniques of collective inference and stacked models, that leverages the homophily demonstrated above. This approach used a combination of link structure and similarity to find "neighbors" and rank them. We also presented a new sub-network approach, that uses similarity to select neighbors for the stacked model, which is not limited to the controversy problem domain, and can be used in other areas. Additionally, we constructed a neighbors-only classifier that does not utilize the features of a page

itself but only on its neighbors, and demonstrated that, counter-intuitively, in certain cases it can be as effective as a classifier that relies on the page's own features.

We have publicly released two major data sets for controversy that were constructed for the purposes of this thesis. One data set (see Chapter 4) contains 377 web pages and 1761 Wikipedia articles annotated with regards to controversy, which is the first data set available for the web classification problem, and the largest data set of controversy labels released to date. The data set also includes 3430 annotations of pairs of webpages and Wikipedia articles, regarding whether or not the Wikipedia page is on the same topic as the webpage. Another data set from Twitter (see Chapter 3) contains nearly 100 million tweets for several popular topics in the last eighteen months, including three prominent controversies (the 2016 U.S. Elections, the UK referendum on leaving the EU, commonly known as Brexit, and "The Dress", a photo that went viral when people disagreed on its colors) and the manually curated hashtags used for stances.

The nascent field of controversy analysis and detection has only emerged over the past several years. As with any new field, the number of questions raised by this thesis is far greater than the answers we have found so far. We end this thesis with a discussion, exploring the social and ethical implications of addressing controversial topics from a computational perspective, and outlining multiple open problems and challenges in this emerging field, making conceptual contributions to the nascent study of controversies. In our published work, we discussed civic, ethical and technical challenges in the subfield; and expounded the scope of open problems of interest to academia and industry [Dori-Hacohen and Allan, 2013, Dori-Hacohen and Allan, 2015, Dori-Hacohen et al., 2015]. We share here portions of these contributions, particularly navigating controversy as a complex search task, the problem of defining controversy and the open issues remaining for the field (including ethical and social implications). Portions of this chapter were previously published in our position paper [Dori-Hacohen et al., 2015].

7.1 Navigating Controversy as a Complex Search Task

As mentioned in Section 1.1, the rise of personalization has created a concern regarding "Filter Bubbles", that is, exposure to a narrower range of viewpoints [Pariser, 2011]. Navigating controversy is thus an increasingly challenging task for search engine users and administrators alike. On one hand, by presenting direct answers to a user's information need, search engines feed into confirmation bias and assist users to remain in their own echo chambers. On the other hand, highlighting a controversy outright may have unintended consequences. The subtle differences between fact disputes and their interpretations, between scientific debates and moral stands, further exacerbate these challenges.

Information has a clear effect on the choices people make. The introduction of Fox News, a channel with clear political leanings, was associated with a shift of 3-8% in voting patterns in presidential elections from 1996 to 2000 towards the channel's opinions [DellaVigna and Kaplan, 2007]. In the health domain, queries about celebrities perceived as anorexic were shown to induce queries indicative of eating disorders [Yom-Tov and Boyd, 2014].

When a user's information need pertains to a controversial topic, their search task becomes complex, as does the process of presenting the "correct" information. Since search engines match keywords to the retrieved documents, users are often left on their own to find the language used to describe different stances of an argument, in order to issue queries to retrieve information about them, and to classify the returned documents into these different views. Should search engines help users explicitly in this process? Should search engines make users aware of the different aspects of a topic or, alternatively, downweight some views (though this may arguably be viewed as censorship)? One way or another, helping the user navigate the controversial topic, along with its different opinions and stances, is a crucial part of the search engine's role in the case of these complex search tasks, be it implicitly or explicitly.

Some might argue that the search engine's role in the case of controversial topics ends at presenting the results in a simple keyword-based "list of ten links" on a Search Engine Results Page (SERP), and that the search engine has no place to take a moral stand. Even presenting the controversy and the various stances on it may not be a simple choice: if search engines explicate the different stances regarding a topic (e.g. presenting pro-anorexia opinions alongside anorexia treatments), this information may nudge people towards harmful behavior, either by exposing them to wrong or harmful information, or because users may stop perceiving search engines as honest brokers of information.

On the other hand, simply providing every result available with no qualification can also be harmful, as disputed claims are allowed to proliferate without any warning to the unsuspecting user. For example, unproven, "quack" medical treatments often put users at risk by warning them not to heed their doctors [American Cancer Society, 2012] [Barrett and Herbert, 2014]. With unfounded claims widespread on the web, there are subtle ethical concerns with settling for a "buyer beware" ("caveat emptor") approach; similar concerns have been raised in the medical realm [Caplan and Levine, 2010]. With concerns of life and death on the balance (e.g., in the case of medical controversies), we should not underestimate the impact of such choices on search engine users. Recent work assumes that trustworthiness should be preserved, for example in the case of knowledge extraction [Dong et al., 2015]. Some may go as far as arguing that, if technology allows for discernment of trustworthy vs. non-trustworthy sources, the search engine has an obligation to serve the trustworthy results to the users; others may say this is a slippery slope, and may in fact be viewed as censorship.

Beyond the complex task that the user herself is trying to complete, complexity also stems from the search engine's design and algorithmic choices. It's possible that amidst all the websites crawled by an engine, the correct response (if one even exists) is nowhere to be found, or is unfairly biased [White and Hassan, 2014]. Should a search engine operator be concerned with civic or ethical implications of the search results it serves on controversial

topics [Hinman, 2005]? Should the user always be provided with what they want to see, even if it can be harmful to the user, or to society as a whole? Where should we draw the line between presenting trustworthy information from authoritative sources and discounting incorrect statements, versus presenting opinions on a moral debate?

These questions are open problems. Far from providing the community with a "correct" answer, we open the discussion on the case of navigating controversy as a complex search task. In our work, we have highlighted some of the issues that users may want to perform when searching for information on controversial topics, including seeking information on controversial topics; understanding different stances or opinions on such topics; and placing results within the context of the larger debate. Even the definition of controversy is still an open question, one which we have advanced in Chapter 3.

7.2 Single truth or shades of gray

Information needs vary in the number of answers to them, both correct and incorrect. Some information needs have a single correct answer to them, while others may have several possible correct answers, requiring a moral judgment or entailing an opinion, e.g. political and religious questions. There are also questions for which there is a single scientifically correct answer, but for which non-scientific responses exist, even though they are factually incorrect. For example, some people claim that the Mumps-Measles-Rubella (MMR) vaccine causes autism; though studies have shown this claim to be incorrect, it is still believed by many people. This variation in answers requires different treatment in each case. The simplest category is that where the information need has a single, correct, answer, which the search engine can provide. The second category is of questions which have a technically correct response, but also an incorrect one which is prevalent on the web. Recent research by White and Hassan has demonstrated this phenomena in web search results, and specifically in health search [White and Hassan, 2014].

The last category is of questions which have several possible correct answers, among which people may choose by making a moral judgment, for example, topics of abortion, same-sex marriage, and other highly charged issues; religious and political questions often fall under this umbrella. Selective exposure theory shows that people seek information which affirms their viewpoint and avoid information which challenges it [Frey, 1986]. Exposure to differing viewpoints has been shown to be socially advantageous in reducing the likelihood of adopting polarized views [Stinchcombe, 2010] and increasing tolerance for people with other opinions [Garrett and Resnick, 2011]; thus, some researchers argue that technology could be used to expose people to a broader variety of perspectives, nudging people to becoming "open-minded deliberators" [Garrett and Resnick, 2011]. This reasoning has led researchers to try and inform people of the differing views on the topics which they are reading. It is technically possible to provide people with diverse opinions where they have sought only one (cf. [Munson et al., 2013, Kriplean et al., 2012, Oh et al., 2009, Yom-Tov et al., 2013]), but there still remains the question of whether a search engine should do so.

An additional concern is whether claiming that certain facts are "true" or "false" holds any objective meaning. The scope of this thesis does not allow a deep dive into the philosophical questions of objectivism vs. moral relativism, and the constructs of objectivity, subjectivity and intersubjectivity (for an exploration of these concepts with regard to stance taking, see Du Bois [Du Bois, 2007]). Nonetheless, we can still delineate a few obvious concerns: the choice of which facts are in dispute, or which topics are controversial, can vary significantly with the cultural and social setting in which these questions are evaluated. For example, a user in Israel and a user in Iran may have very different opinions about what holds "true", and either may be offended if the others' worldview was presented as a "fact"; what is fact to one is either highly controversial or simply false to the other, and vice versa. As another example, the research by White and Hassan cited above [White and Hassan, 2014] assumes that the Western world's view of medicine is the only correct one,

but users in China may beg to differ. Is a topic therefore only controversial if a user (or culture) believes it to be so? Who, then, can decide when a topic is controversial? How can the system know that a user believes a topic is controversial, and should the system then respond differently than when a user accepts it as "fact"? As of this writing, concerns regarding fake news and so-called "alternative facts" have been raised to national and international consciousness, which serves to reinforce these questions which we first posed in 2015 [Dori-Hacohen et al., 2015].

7.3 Open Questions

Search engines are claimed to hold significant political power [Introna and Nissenbaum, 2000], and deliberative democracy's basic tenets is arguably the ability to have a shared set of experiences, and to be exposed to arguments you disagree with [Sunstein, 2009]. Search engines and social media are increasingly responsible for "Filter Bubbles", wherein click-feedback and personalization lead users to only see what they want, serving to further increase confirmation bias [Pariser, 2011]. While this may seems to match individual users' preference, the net effect on society is potentially detrimental. Being exposed only to like-minded people in so-called "echo chambers" serves to increase polarization and reduce diversity [Schkade et al., 2007][1].

Contrary to the common wisdom, some evidence exists that online personalization has not increased the filter bubble [Gentzkow and Shapiro, 2011]. That said, research has shown that exposing users to opposing opinions increases their interest in seeking diverse opinions, and their interest in news in general [Yom-Tov et al., 2013]. There have been suggestions to diversify search results based on sentiment [Kacimi and Gamper, 2012], though others argue that presenting the opposite opinion would only help in some cases [An et al., 2013, Munson and Resnick, 2010]. Prior bias of people changes the results of a search

[1] We note that, despite our own biases, the values of democracy and diversity of opinion are also culturally predicated, and not necessarily applicable to all search engine users.

query, even without personalization. For example, the results for the query "what are the advantages of the MMR vaccine?" are completely different from the results served for the query "what are the dangers of the MMR vaccine?". Moreover, the way people interpret the same information is dependent on their bias, for example in the case of gun control [Koutra et al., 2014] or bias towards vaccines [Yom-Tov et al., 2014]. Thus, if a user seeks information on "how does MMR cause autism?", should a search engine inform the user of the truth, or just satisfy their information need? One possible solution includes highlighting disputed claims [Ennals et al., 2010] or explicitly presenting opposing viewpoints [Vydiswaran et al., 2012], but the problem remains that the user may not trust sources that don't match their existing worldview.

Since search engines (as well as their social media counterparts) are increasingly the dominant medium for seeking information and news, the question then becomes: should search engines reflect what is on the internet and match content to users to maximize their preference, regardless of its truth value, or any concerns about diversity of opinion? Where do we draw the line between fact disputes and moral debates? Should the controversial nature of a topic depend on the social and cultural setting in which it is being evaluated? Should search engines have a civic duty, and in that case, who decides what that duty is?

There are multiple technical challenges remaining in classifying controversial topics and extracting the opinions about them; we address some of them in this work, though many others remain beyond the scope of this thesis. We see the controversy detection problem (see Chapters 4 & 5) as a prerequisite to several other interesting applications and larger problems such as: user studies on the effects of informing users when the webpage they are looking at is controversial; the evolution and incidence of controversial topics over time; and diversifying controversial search results according to the stances on them, are a few such problems. However, even if these technical challenges of detecting controversy and stances were solved, there remains the question of if, when and how to present these to the user, based on their information need. As we discussed, there are ethical concerns with

a search engine taking action, but also with inaction. It remains to be seen if users would be interested in hearing opposing opinions, or whether interventions would be useful; and finally, it is unclear whether it is within the search engine's purview (or even its duty) to intervene, and if so, how. By introducing these questions in a clear and systematic way into the research community, both through our technical work about the definition of controversy (Chapter 3) and through our positions outlined in this chapter, we significantly contributed to the theoretical and conceptual understanding of this emerging field, and hope to spark further discussions that will positively impact the direction of future research by others.

In addition to these open questions, several larger areas of study remain. For example, how does controversy emerge? How does it diffuse in society, both offline and online, in media and in social networks? What is the effect of the filter bubble as well as offline or self-chosen echo chambers? What are the quantitative and qualitative differences between different types of controversies, such as political, moral, scientific, medical or religious? How does controversy relate to the well-known challenge of fact disputes [Ennals et al., 2010] and the more recent rash of fake news? Elaborating on the distinctions between facts, beliefs and conspiracy theories [Jolley and Douglas, 2014, Bessi et al., 2015], especially in this era dominated by a discussion of fake news [Shao et al., 2016], will allow further understanding of misinformation and disinformation, which are growing becoming increasingly intertwined with controversy. Search engines and online social networks such as Google and Facebook are starting to become aware of these challenges, as is the general population, but potential solutions are still few and far between. These are all questions far outside the scope of this thesis, and will provide a rich basis for many research projects for years to come. Likewise, they are issues impacting billions of users' lives, and they are by no means solved. These fascinating questions are out of scope for this paper; we look forward to exploring them in future work, and invite others to join us.

Given the state of the field when embarking on this thesis - with only a small handful of papers exploring these questions prior to 2011 - it becomes clear that we have made a significant contribution to the state of the art in the field. At the same time, we currently see scope for much greater inquiry arising from the research performed so far. Likewise, given the interdisciplinary nature of controversies, a greater need for collaboration among various branches of humanities and social sciences as well as various subfields in computer science is clearly called for.

BIBLIOGRAPHY

[Abiteboul et al., 2000] Abiteboul, S., Buneman, P., and Suciu, D. (2000). *Data on the Web: from relations to semistructured data and XML.* Morgan Kaufmann.

[Aktolga and Allan, 2013] Aktolga, E. and Allan, J. (2013). Sentiment Diversification With Different Biases. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 593–602.

[Amendola et al., 2015] Amendola, L., Marra, V., and Quartin, M. (2015). The evolving perception of controversial movies. *Palgrave Communications*, 1.

[American Cancer Society, 2012] American Cancer Society (2012). Metabolic Therapy.

[An et al., 2013] An, J., Quercia, D., and Crowcroft, J. (2013). Why individuals seek diverse opinions (or why they don't). *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*, pages 15–18.

[Awadallah et al., 2011] Awadallah, R., Ramanath, M., and Weikum, G. (2011). OpinioNetIt: understanding the opinions-people network for politically controversial topics. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2481–2484, New York, NY, USA. ACM.

[Awadallah et al., 2012a] Awadallah, R., Ramanath, M., and Weikum, G. (2012a). Harmony and Dissonance : Organizing the People's Voices on Political Controversies. *New York*, pages 523–532.

[Awadallah et al., 2012b] Awadallah, R., Ramanath, M., and Weikum, G. (2012b). Harmony and Dissonance: Organizing the People's Voices on Political Controversies. *WSDM*, pages 523–532.

[Barrett and Herbert, 2014] Barrett, S. and Herbert, V. (2014). Twenty-Six Ways to Spot Quacks and Vitamin Pushers.

[Bessi et al., 2015] Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., and Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PLoS ONE*, 10(2):1–17.

[Bex et al., 2014] Bex, F., Snaith, M., Lawrence, J., and Reed, C. (2014). ArguBlogging: An application for the Argument Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 25:9–15.

[Borra et al., 2015] Borra, E., Kaltenbrunner, A., Mauri, M., Amsterdam, U., Weltevrede, E., Laniado, D., Rogers, R., Ciuccarelli, P., and Magni, G. (2015). Societal Controversies in Wikipedia Articles. *Proceedings CHI 2015*, pages 3–6.

[Brandes et al., 2009] Brandes, U., Kenis, P., Lerner, J., and van Raaij, D. (2009). Network analysis of collaboration structure in Wikipedia. *Proceedings of the 18th international conference on World wide web - WWW '09*, page 731.

[Brandes and Lerner, 2008] Brandes, U. and Lerner, J. (2008). Visual Analysis of Controversy in User-Generated Encyclopedias. *Information Visualization*, 7(1):443–452.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Bröcheler et al., 2012] Bröcheler, M., Mihalkova, L., and Getoor, L. (2012). Probabilistic Similarity Logic. *CoRR*, abs/1203.3.

[Bykau et al., 2015] Bykau, S., Korn, F., Srivastava, D., and Velegrakis, Y. (2015). Fine-Grained Controversy Detection in Wikipedia.

[Callan et al., 1992] Callan, J. P., Croft, W. B., and Harding, S. M. (1992). The {INQUERY} retrieval system. In *Database and Expert Systems Applications*, pages 78–83. Springer Vienna.

[Caplan and Levine, 2010] Caplan, A. and Levine, B. (2010). Hope, hype and help: Ethically assessing the growing market in stem cell therapies. *Current*, 10(5):33–34.

[Cartright et al., 2009] Cartright, M.-A., Aktolga, E., and Dalton, J. (2009). Characterizing the Subjectivity of Topics. In *SIGIR*. Springer.

[Cedroni, 2010] Cedroni, L. (2010). Voting Advice Applications in Europe: A Comparison. *Voting Advice Applications in Europe: The State of Art*, pages 247–258.

[Chakrabarti et al., 1998] Chakrabarti, S., Dom, B., and Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. *ACM SIGMOD Record*, 27(2):307–318.

[Chen and Berger, 2013] Chen, Z. and Berger, J. (2013). When, Why, and How Controversy Causes Conversation. *Journal of Consumer Research*, 40(3):580–593.

[Cheng and Roth, 2013] Cheng, X. and Roth, D. (2013). Relational inference for wikification. *Urbana*.

[Choi et al., 2010] Choi, Y., Jung, Y., and Myaeng, S.-H. (2010). Identifying Controversial Issues and Their Sub-topics in News Articles. *Intelligence and Security Informatics*, 6122:140–153.

[Coletto et al., 2016] Coletto, M., Lucchese, C., Orlando, S., and Perego, R. (2016). Polarized user and topic tracking in twitter. In *SIGIR*, pages 945–948. ACM.

[Cramer, 2011] Cramer, P. A. (2011). *Controversy as News Discourse*. Argumentation Library. Springer Netherlands.

[Das et al., 2013] Das, S., Lavoie, A., and Magdon-Ismail, M. (2013). Manipulation Among the Arbiters of Collective Intelligence: How Wikipedia Administrators Mold Public Opinion. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 1097–1106, New York, NY, USA. ACM.

[Dascal, 1995] Dascal, M. (1995). Epistemology, Controversies, and Pragmatics. *Isegoría*, 12(8-43).

[De Choudhury et al., 2016] De Choudhury, M., Jhaver, S., Sugar, B., and Weber, I. (2016). Social Media Participation in an Activist Movement for Racial Equality. In *Tenth International AAAI Conference on Web and Social Media*, pages 92–101.

[DellaVigna and Kaplan, 2007] DellaVigna, S. and Kaplan, E. (2007). The Fox News effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234.

[Dietz et al., 2007] Dietz, L., Bickel, S., and Scheffer, T. (2007). Unsupervised prediction of citation influences. *Proceedings of the 24th International Conference on Machine Learning (2007)*, 227:233–240.

[Dong et al., 2015] Dong, X. L., Gabrilovich, E., Murphy, K., Dang, V., Watts, I., Horn, W., Lugaresi, C., Sun, S., and Zhang, W. (2015). Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. *Arxiv preprint*, (Section 3).

[Dori-Hacohen and Shavit, 2013] Dori-Hacohen, G. and Shavit, N. (2013). The cultural meanings of Israeli Tokbek (talk-back online commenting) and their relevance to the online democratic public sphere. *International Journal of Electronic Governance*, 6(4):361–379.

[Dori-Hacohen and Allan, 2013] Dori-Hacohen, S. and Allan, J. (2013). Detecting controversy on the web. In *Proceedings of the 22nd ACM international conference on Conference on Information & Knowledge Management*, CIKM '13, pages 1845–1848, New York, NY, USA. ACM.

[Dori-Hacohen and Allan, 2015] Dori-Hacohen, S. and Allan, J. (2015). Automated Controversy Detection on the Web. In *Proceedings of the 37th European Conference on Information Retrieval (ECIR 2015)*, pages 423–434.

[Dori-Hacohen et al., 2015] Dori-Hacohen, S., Yom-Tov, E., and Allan, J. (2015). Navigating Controversy as a Complex Search Task. In *Proceedings of the first international workshop on Supporting Complex Search Tasks, volume 1338 of CEUR Workshop Proceedings*.

[Du Bois, 2007] Du Bois, J. W. (2007). The stance triangle. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, pages 139–182.

[Ennals et al., 2010] Ennals, R., Trushkowsky, B., and Agosta, J. M. (2010). Highlighting disputed claims on the web. In *Proceedings of the 19th international conference on World wide web - WWW '10*, WWW '10, page 341, New York, New York, USA. ACM Press.

[Fast and Jensen, 2008] Fast, A. and Jensen, D. (2008). Why stacked models perform effective collective classification. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 785–790.

[Fogg et al., 2003] Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., and Tauber, E. R. (2003). How do users evaluate the credibility of Web sites?: a study with over 2,500 participants. *Proceedings of the 2003 conference on Designing for user experiences*, pages 1–15.

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

[Garimella et al., 2016] Garimella, K., Morales, G. D. F., Gionis, A., and Mathioudakis, M. (2016). Quantifying controversy in social media. *WSDM*, pages 1–10.

[Garrett and Resnick, 2011] Garrett, R. K. and Resnick, P. (2011). Resisting political fragmentation on the Internet. *Daedalus*, 140(4):108–120.

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

[Goldsmith-Pinkham and Imbens, 2013] Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social Networks and the Identification of Peer Effects. *Journal of Business & Economic Statistics*, 31(3):253–264.

[Gyllstrom and Moens, 2011] Gyllstrom, K. and Moens, M.-F. M. (2011). Clash of the typings: finding controversies and children's topics within queries. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 80–91, Berlin, Heidelberg. Springer.

[Hasan and Ng, 2013] Hasan, K. S. K. and Ng, V. (2013). Frame Semantics for Stance Classification. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, page 124.

[Helfand, 2016] Helfand, D. J. (2016). *A Survival Guide to the Misinformation Age: Scientific Habits of Mind*. Columbia University Press.

[Heroic Media, 2014] Heroic Media (2014). Free Abortion Help website.

[Hinman, 2005] Hinman, L. M. (2005). Esse est indicato in Google: Ethical and political issues in search engines. *International Review of Information Ethics*, 3(6):19–25.

[Holderness, 2015] Holderness, C. (2015). What Colors Are This Dress? https://www.buzzfeed.com/catesish/help-am-i-going-insane -its-definitely-blue, accessed: 2017-01-13.

[Introna and Nissenbaum, 2000] Introna, L. D. and Nissenbaum, H. (2000). Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society*, 16(3):169–185.

[Jang et al., 2016] Jang, M.-h., Foley, J., Dori-Hacohen, S., and Allan, J. (2016). Probabilistic Approaches to Controversy Detection. In *CIKM*.

[Jankowski-Lorek et al., 2014] Jankowski-Lorek, M., Nielek, R., Wierzbicki, A., and Kazimierz Zielinski (2014). Predicting Controversy of Wikipedia Articles Using the Article Feedback Tool.

[Jensen et al., 2004] Jensen, D., Neville, J., and Gallagher, B. (2004). Why collective inference improves relational classification. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, page 593.

[Jesus et al., 2009] Jesus, R., Schwartz, M., and Lehmann, S. (2009). Bipartite networks of Wikipedia's articles and authors: a meso-level approach. *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*.

[Jolley and Douglas, 2014] Jolley, D. and Douglas, K. M. (2014). The effects of anti-vaccine conspiracy theories on vaccination intentions. *PLoS ONE*, 9(2):e89177.

[Journal of Vision Special Collection, 2016] Journal of Vision Special Collection (2016). A Dress Rehearsal for Vision Science.

[Kacimi and Gamper, 2012] Kacimi, M. and Gamper, J. (2012). MOUNA: Mining Opinions to Unveil Neglected Arguments. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, page 2722, New York, New York, USA. ACM Press.

[Kahan, 2015] Kahan, D. M. (2015). Climate-science communication and the measurement problem. *Political Psychology*, 36(S1):1–43.

[Kittur et al., 2007] Kittur, A., Suh, B., Pendleton, B. A., Chi, E. H., Angeles, L., and Alto, P. (2007). He Says, She Says: Conflict and Coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 453–462, New York, NY, USA. ACM Press.

[Klenner et al., 2014] Klenner, M., Amsler, M., and Hollenstein, N. (2014). Verb Polarity Frames: a New Resource and its Application in Target-specific Polarity Classification. In *Proceedings of the 12th edition of the KONVENS conference Vol. 1. - Hildesheim*. Universität Hildesheim.

[Kou and Cohen, 2007] Kou, Z. and Cohen, W. W. (2007). Stacked Graphical Models for Efficient Inference in Markov Random Fields. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 533–538.

[Koutra et al., 2014] Koutra, D., Bennett, P., and Horvitz, E. (2014). Events and Controversies: Influences of a Shocking News Event on Information Seeking. *TAIA workshop in SIGIR*, pages 0–3.

[Kriplean et al., 2012] Kriplean, T., Morgan, J., Freelon, D., Borning, A., and Bennett, L. (2012). Supporting reflective public thought with ConsiderIt. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 265–274. ACM.

[Kuwadekar and Neville, 2011] Kuwadekar, A. and Neville, J. (2011). Relational active learning for joint collective classification models. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 385–392.

[Leibniz, 1982] Leibniz, G. W. (1982). Vorausedition zur Reihe VI (Philosophische Schriften) in der Ausgabe der Akademie Wissenschaften der DDR. *M{ü}nster: Leibniz-Forschungsstelle der Universit{ä}t M{ü}nster*, 1991:1253.

[Leshner, 2015] Leshner, A. I. (2015). Bridging the opinion gap. *Science*, 347(6221):459.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

[McDonald, 2017] McDonald, M. P. (2017). 2016 November General Election Turnout Rates, united states elections project. http://www.electproject.org/2016g, accessed 2017-01-12.

[McPherson et al., 2001] McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27:pp. 415–444.

[Mejova et al., 2014] Mejova, Y., Zhang, A. X., Diakopoulos, N., and Castillo, C. (2014). Controversy and Sentiment in Online News.

[Milajevs and Bouma, 2013] Milajevs, D. and Bouma, G. (2013). Real time discussion retrieval from twitter. In *WWW*.

[Munson et al., 2013] Munson, S. A., Lee, S. Y., and Resnick, P. (2013). Encouraging Reading of Diverse Political Viewpoints with a Browser Widget. In *Proceedings of the Internation Conference on Weblogs and Social Media*.

[Munson and Resnick, 2010] Munson, S. A. and Resnick, P. (2010). Presenting Diverse Political Opinions: How and How Much. In *Proc. CHI 2010*, CHI '10, pages 1457–1466, New York, NY, USA. ACM.

[Oakley and Berlin, 1946] Oakley, A. and Berlin, I. (1946). *Annie Get Your Gun (Musical)*.

[Oh et al., 2009] Oh, A., Lee, H., and Kim, Y. (2009). User evaluation of a system for classifying and displaying political viewpoints of weblogs. *Proc. ICWSM*.

[Pariser, 2011] Pariser, E. (2011). *The Filter Bubble: What the Internet is hiding from you*. Penguin Press HC.

[Pew Research Center, 2015a] Pew Research Center (2015a). An Elaboration of AAAS Scientists' Views. Technical report.

[Pew Research Center, 2015b] Pew Research Center (2015b). Public and Scientists' Views on Science and Society. Technical report.

[Popescu and Pennacchiotti, 2010] Popescu, A.-M. and Pennacchiotti, M. (2010). Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, pages 1873–1876.

[Riedel et al., 2010] Riedel, S., Yao, L., and McCallum, A. (2010). Modeling Relations and Their Mentions Without Labeled Text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, ECML PKDD'10, pages 148–163, Berlin, Heidelberg. Springer-Verlag.

[Rijsbergen, 1979] Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

[Rogers, 2015] Rogers, A. (2015). The Science of Why No One Agrees on the Color of This Dress.

[Rose et al.,] Rose, T., Stevenson, M., and Whitehead, M. The Reuters Corpus Volume 1. In *LREC*.

[Schkade et al., 2007] Schkade, D., Sunstein, C. R., and Hastie, R. (2007). What happened on deliberation day? *California Law Review*, 95(298):915–940.

[Schlaffke et al., 2015] Schlaffke, L., Golisch, A., Haag, L. M., Lenz, M., Heba, S., Lissek, S., Schmidt-Wilcke, T., Eysel, U. T., and Tegenthoff, M. (2015). The brain's dress code: How The Dress allows to decode the neuronal pathway of an optical illusion. *Cortex*, 73:271–275.

[Sepehri Rad and Barbosa, 2012] Sepehri Rad, H. and Barbosa, D. (2012). Identifying controversial articles in Wikipedia: A comparative study. In *Proceedings of 8th conference on WikiSym*, WikiSym '12. ACM.

[Sepehri Rad et al., 2012] Sepehri Rad, H., Makazhanov, A., Rafiei, D., and Barbosa, D. (2012). Leveraging editor collaboration patterns in Wikipedia. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, HT '12, page 13, New York, New York, USA. ACM Press.

[Shao et al., 2016] Shao, C., Ciampaglia, G. L., Flammini, A., and Menczer, F. (2016). Hoaxy: A Platform for Tracking Online Misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 745–750. International World Wide Web Conferences Steering Committee.

[Stinchcombe, 2010] Stinchcombe, A. L. (2010). Going to Extremes: How Like Minds Unite and Divide. *Contemporary Sociology: A Journal of Reviews*, 39(2):205–206.

[Sumi et al., 2011] Sumi, R. R., Yasseri, T., Rung, A., Kornai, A., and Kertész, J. (2011). Edit wars in Wikipedia. *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on Social Computing (SocialCom)*, pages 724–727.

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

[Szívós, 2005] Szívós, M. (2005). Temporality, reification and subjectivity. *Controversies and Subjectivity*, 1:201.

[The Electoral Comission, 2016] The Electoral Comission (2016). EU referendum results. http://www.electoralcommission.org.uk/find-information-by-subject/elections-and-referendums/past-elections-and-referendums/eu-referendum/electorate-and-count-information, accessed: 2017-01-12.

[Tsytsarau et al., 2011] Tsytsarau, M., Palpanas, T., and Denecke, K. (2011). Scalable detection of sentiment-based contradictions. *DiversiWeb 2011*.

[Van Eemeren and Garssen, 2008] Van Eemeren, F. H. and Garssen, B. (2008). *Controversy and confrontation: Relating controversy analysis with argumentation theory*, volume 6. John Benjamins Publishing.

[Viégas et al., 2004] Viégas, F. B., Wattenberg, M., and Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, 6(1):575–582.

[Vuong et al., 2008] Vuong, B.-q., Lim, E.-p., Sun, A., Le, M.-T., Lauw, H. W., and Chang, K. (2008). On ranking controversies in Wikipedia: models and evaluation. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 171–182, New York, NY, USA. ACM.

[Vydiswaran et al., 2012] Vydiswaran, V. G. V., Zhai, C., Roth, D., and Pirolli, P. (2012). BiasTrust: Teaching Biased Users About Controversial Topics. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1905–1909, New York, NY, USA. ACM.

[Walia, 2013] Walia, A. (2013). 22 Medical Studies That Show Vaccines Can Cause Autism.

[Wang and Cardie, 2014] Wang, L. and Cardie, C. (2014). A Piece of My Mind : A Sentiment Analysis Approach for Online Dispute Detection. *ACL*, pages 693–699.

[Wang and Sukthankar, 2013] Wang, X. and Sukthankar, G. (2013). Multi-label Relational Neighbor Classification Using Social Context Features. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 464–472, New York, NY, USA. ACM.

[Wasserman, 2017] Wasserman, D. (2017). 2016 National Popular Vote Tracker. http://cookpolitical.com/story/10174, accessed: 2017-01-12.

[White and Hassan, 2014] White, R. W. and Hassan, A. (2014). Content bias in online health search. *ACM Transactions on the Web (TWEB)*, 8(4):25.

[Wikipedia, 2014] Wikipedia (2014). Wikipedia: {Neutral} {Point} of {View} Policy.

[Wikipedia, 2016] Wikipedia (2016). Toilet paper orientation. https://en.wikipedia.org/wiki/Toilet_paper_orientation, accessed: 2016-10-23.

[Wikipedia, 2017] Wikipedia (2017). United States presidential election, 2016. https://en.wikipedia.org/wiki/United_States_presidential_election,_2016, date accessed: 2017-01-13.

[Wilson and Likens, 2015] Wilson, A. M. and Likens, G. E. (2015). Content volatility of scientific topics in Wikipedia: A cautionary tale. *PLoS ONE*, 10(8):10–14.

[Yasseri et al., 2014] Yasseri, T., Spoerri, A., Graham, M., and Kertész, J. (2014). The most controversial topics in Wikipedia: A multilingual and geographical analysis. In *Global Wikipedia: International and cross-cultural issues in collaboration*, page 178.

[Yasseri et al., 2012] Yasseri, T., Sumi, R., Rung, A., Kornai, A., and Kertész, J. (2012). Dynamics of conflicts in Wikipedia. *PloS one*, 7(6):e38869.

[Yom-Tov and Boyd, 2014] Yom-Tov, E. and Boyd, d. m. (2014). On the link between media coverage of anorexia and pro-anorexic practices on the web. *International Journal of Eating Disorders*, 47(2):196–202.

[Yom-Tov et al., 2013] Yom-Tov, E., Dumais, S. T., and Guo, Q. (2013). Promoting civil discourse through search engine diversity. *Social Science Computer Review*.

[Yom-Tov et al., 2014] Yom-Tov, E., Fernandez-Luque, L., and Luque, L. (2014). Information is in the eye of the beholder: Seeking information on the {MMR} vaccine through an Internet search engine. In *Proceedings of the American Medical Informatics Association*.