Modeling Controversy within Populations

Myungha Jang, Shiri Dori-Hacohen and James Allan
Center for Intelligent Information Retrieval
College of Information and Computer Sciences
University of Massachusetts

ABSTRACT

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. Since controversy is a complicated social phenomenon, it is difficult to understand what elements make up the controversy. Previous work has attempted to capture controversy algorithmically by studying cues for disagreement and polarity between different stance groups. However, we still lack a systematic understanding of how controversy should be defined and measured. In this paper, we propose a multi-dimensional model of controversy. Specifically, we introduce a model with two minimal dimensions: contention and importance. Our model departs from existing work by viewing controversy as a trait rooted in population. It suggests that controversy should be separately observed in a given population, rather than held as a fixed universal quantity. We model contention and importance within a population from a mathematical standpoint. To validate and evaluate the soundness of our theoretical model, we instantiate the model to algorithms for a diverse set of sources: polling, Twitter, and Wikipedia. We demonstrate that our controversy model holds an explanatory power for observed phenomena but also a predictive power for tasks, such as identifying controversial Wikipedia articles.

然而，我们对如何解决争议，如何表现以及如何衡量它的理论和实践理解仍存在差距

争论，重要程度。
争议应该在特定人群中单独观察，而不是作为一个固定的普遍数量。

我们证明了我们的争议模型对观察到的现象具有解释力，但也对任务具有预测能力

1 INTRODUCTION

Social networks, such as Twitter, Facebook, and discussion boards, have become one of the most popular places where controversial arguments are held. Accordingly, technological tools have also become critical in shaping these discussions by curating and filtering the content seen by each user. From a computational perspective, we currently do not understand controversy well enough. Algorithms based on an incomplete understanding of controversy are bound to fail in unexpected ways, which can replicate or even exacerbate the sources of human bias.

Recent work on controversy cuts across traditional disciplinary lines—including a wide variety of computational tasks along with the social sciences and the humanities—and has made significant strides in analyzing and detecting controversy (cf. [3, 12]). Nevertheless, there are serious gaps in our theoretical and practical understanding of how to define controversy, how it manifests, and how it evolves. For example, polling organizations naturally select topics of broad interest and segment their results based on certain populations defined by demographics such as race and gender. These notions are surprisingly absent from algorithmic analyses of controversy. Instead, controversy has only been defined as an absolute, single value for an amorphous global population.

Meanwhile, there has been a growing disparity between scientific understanding and public opinion on certain controversial topics, such as climate change, evolution, and vaccination [19], with many scientists fighting these trends by arguing that "there is no controversy" [13] (referring to *scientific* controversy). Still, non-scientific claims and arguments have continued to proliferate, raising exposure to the supposedly non-existent controversies. As researchers studying controversies online, how can we reconcile the oft-repeated argument from the scientific community that "there is no controversy" with the practical appearance of wildly diverse opinions on many topics? In other words, is a topic like climate change controversial?[1]

该理论模型将争议视为至少“争议”和“重要性”的组合，与特定人群相关

We address these issues by proposing a theoretical model that defines controversy as a combination of at least "contention" and "importance" with respect to a given population. The model therefore captures the idea that not all controversies are of equal interest. It also suggests that the right question to be asked is not "is climate change controversial?" but instead "is climate change controversial to *a particular group*?"

和之前工作的不同

Our framework departs from existing work on controversy in several important ways. First, we define controversy in terms of not only its topic but also a given population. Second, our model accounts for participants in the population who hold no stance on a specific topic and also allows for any number of stances rather than just a strict dichotomy. Third, our model allows that some items may be less controversial because they are contentious but not important (and vice versa). These elements give our model explanatory power that can be used to understand a large variety of observed phenomena, ranging from international conflict and community-specific controversies to high-stakes public controversies on well-understood phenomena such as climate change, evolution, and vaccination.

使用真实数据集

In order to ground our theoretical model, we examine a diverse collection of datasets from both online and offline sources. First, we examine several real-world polling datasets, including a poll that focuses on opinions about scientific topics, such as climate change and evolution, measured among the general US population and the scientific community [20]. Additionally, we look at Twitter coverage for three prominent controversies: the 2016 US Elections, the UK European Union membership referendum, commonly known

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

ICTIR'17, October 1–4, 2017, Amsterdam, The Netherlands.
 ISBN 978-1-4503-4490-6/17/10…$15.00
DOI: http://dx.doi.org/10.1145/3121050.3121067

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," real voter data for Brexit, and the US Elections. Finally, we apply our model to Wikipedia and show that it can predict whether or not a Wikipedia article is controversial.

2 RELATED WORK

Most recent work on controversy has measured controversy as either a binary state or a single quantity, both of which are to be measured or estimated directly [2, 3, 15, 22]. With a few exceptions [1, 16], earlier work did not model controversy formally. Even when it did, the meaning of controversy was not modeled, but instead assumed to be a known quantity in the world. Most prior work in computer science does not define controversy and treats it as a global quantity (cf.[17, 28]). Past research shows that achieving inter-annotator agreement on the "controversy" label is challenging [10, 18].

Meanwhile, most of the work on controversy in the social sciences and the humanities is qualitative by nature and often focuses on one or two examples of controversy (c.f. [11, 24]). Otherwise, it works toward a more qualitative analysis of the overall patterns across controversies [9] with one notable exception [8]. 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" [6]. However, these definitions leave a gap when people disagree on opinions that are strongly held on less important topics such as the color of a dress, the orientation of toilet paper. There may be "opposing viewpoints" in these topics, but there is no real controversy.

We depart from past research by modeling controversy as a multi-dimensional quantity within population. Similarly, Timmermans et al. also identified five aspects of controversy in news articles: time persistence, emotion, multitude of actors, polarity, and openness. While their approach is closest to our work, they have specifically targeted news articles and have not focused on actual modeling [26].

3 MODELING POPULATION-BASED CONTROVERSY

Controversy is a complicated social phenomenon. Since it is difficult to formally and systematically define what controversy is, there has been little effort to formulate a model that quantifies the level of controversy for a given topic.

As a motivating example, consider two controversies: "The Dress" and the Brexit referendum. First, "The Dress" refers to a photo that went viral over social media starting on Feb. 26, 2015 after people could not agree on its colors. The photo was posted to tumblr and made popular by a Buzzfeed article that asked "What color is this dress?" in a poll with two options: black and blue or gold and white. To date, over 37 million people viewed the article [14]. Second, the Brexit referendum, officially known as the UK 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 European Union. The referendum had immediate political and

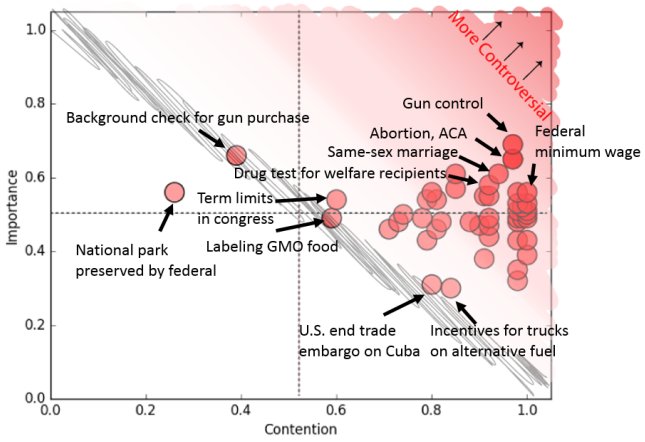

Figure 1: iSideWith topics plotted in the two-dimensional controversy plot of contention and importance. Contention scores are computed by our definition of contention and importance was given by users in the dataset. Sample topics are given in each quadrant of {low,high} importance and contention.

financial ramifications, including the worst one-day drop in the global stock market in history,

When observed by populations interested in each controversy, "The Dress" and Brexit were both extremely contentious. Nearly any group of people sampled from these populations was strongly divided in their opinion. However, it is clear that placing "The Dress" and Brexit in the same bucket is problematic. Brexit affects the fate of entire nations, with far-reaching consequences on diplomatic relationships and the world economy. The other, a photo of the dress, caused a surprising divided reaction in color perception, went viral around the world, and was subsequently forgotten by nearly everyone. In other words, it had little impact on the world.

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*. However, this dimension alone cannot adequately explain less important controversies like "The Dress." An additional orthogonal metric is needed in order to distinguish between contention and controversy. Therefore, we hypothesize the existence of a notion of *importance* as a novel dimension of controversy. While more dimensions can be investigated, we hypothesize that these are the core features of controversy.

This framework is illustrated schematically in Figure 1, overlaying actual results including importance reported in the iSideWith dataset (see Table 2). The first dimension is contention, which we define as 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 reported by users of iSideWith. In Figure 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 5). Figure 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. Therefore, they are more controversial. Issues, such as whether the government should provide incentives for trucks to run on alternative fuels, is highly contentious but less important according to users.

Finally, we model the probability of controversy with a given topic T and a given population Ω. Let *Cont* a binary random variable, which denotes the presence of controversy. Simiarly, let C and I be binary random variables, which denotes the presence of contention and importance of topic T.

We model $P(Cont|\theta)$, where $\theta = \{T, \Omega\}$ as the probability that topic T is controversial within the population Ω. Our model hypothesizes that the probability of controversy given T and Ω is the joint probability of two dimensions: contention (C) and importance (I), to be more rigorously defined later.

$$P(Cont|\theta) = P(C, I|\theta)$$

Here, $P(C, I|\theta)$ can be further decomposed as following:

$$P(C, I|\theta) = \frac{P(C, I, \theta)}{P(\theta)} = \frac{P(I|C, \theta) \cdot P(C|\theta) \cdot P(\theta)}{P(\theta)} = P(I|C, \theta) \cdot P(C|\theta)$$

To compute $P(I|C, \theta)$, the correlation between contention and importance of topic to a population has to be identified. While it is difficult to estimate the exact correlation in the real world, we assume that contention and importance are independent of each other, consisting of orthogonal dimensions of controversy. We therefore let $P(I|C, \theta) = P(I|\theta)$.

$$P(Cont|\theta) = P(C|\theta) \cdot P(I|\theta)$$

We now discuss the modeling of $P(C|\theta)$ and $P(I|\theta)$ and how to estimate them from the real data. Refer to Table 1 for the summary of the notations that are used in the model.

Table 1: Notation summary

Symbol	Definition
Ω	a population
$P(Cont\|\theta)$	the probability that topic T is controversial in a given population Ω
$P(C\|\theta)$	the probability that T is contentious in Ω
$P(I\|\theta)$	the probability that T is important in Ω
C	a binary random variable for contention
c	a binary value to set C to be "contentious"
nc	a binary value to set C to be "non-contentious"
s	a stance with regard to topic T
$holds(p, s, T)$	a person p holds stance s with regard to topic T
$affected(T, p)$	a binary function that returns whether or not a person p is affected by topic T
$\hat{S}$	k stances with regard to topic T
s_0	a lack of stance
G_i	a group of people who hold stance s_i
O_i	an opposing group in the population that hold a stance that conflicts with s_i

3.1 Modeling Contention from Population

We define a measure we call *contention*, 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 to be mutually exclusive.

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 population Ω. We define C to be a binary variable to denote whether or not a given topic is contentious. Let us also define two binary values c and nc for C, each of which respectively means contentious and non-contentious. For example, $P(C = c|\Omega, T)$ denotes the probability that T is contentious in Ω, which we will simply denote it as $P(c|\Omega, T)$. Therefore, $P(c|\Omega, T) + P(nc|\Omega, T) = 1$ satisfies.

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 because 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 us define a measure $conflict$ to denote how much two stances are exclusive. $P(conflict|s_i, s_j) = 1$ indicates that s_i and s_j are in a complete conflict, meaning mutually-exclusive. In other words, $P(conflict|s_i, s_j) = 0$ indicates that two stances completely agree with each other. This probability measures the severity of conflict between two stances.

Not all stances are necessarily mutually exclusive or in a complete agreement. A stance of "abortion should be legalized only in certain circumstances" is not mutually exclusive to any of "pro-choice" or "pro-life" stances. In this case, the probability is somewhere in between 0 and 1. Note that a person can hold multiple stances simultaneously as long as none of the two stances are mutually exclusive. Also, not any stance can be jointly held with s_0. By definition, $P(conflict|s_i, s_i) = 0$ and $P(conflict|s_0, s_i) = 0$ satisfy.

Let a **stance group** in the population be a group 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, $\Omega = \bigcup_i G_i$. We let $P(conflict|G_i, G_j)$ be a probability that two groups of G_i and G_j are in a conflict, similarly as it was defined for two stances. 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:

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

Alternatively:

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

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 $g_i = G_i \cap \omega$. 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 $\omega \subseteq \Omega$,

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

Mutually exclusive stances. Most controversial topics have at least two exclusive mutually stances to bisect the community. In this section, we focus on the case of mutually exclusive stances and describe the model that can be defined. The model described here can be easily generalized by adjusting the value of $P(conflict|s_i,s_j)$.

Intuitively, a group is in the most contentious state when it is equally divided among the two stances. On the other hand, if a group is unequally divided among each stance, the stance of the smaller subgroup would be considered the minority, contributing to a smaller contention score. Our model captures this insight below.

Recall that stance group G_i is defined as the population of people who hold a stance s_i on T. We also define an **opposing group** in the population be a group of people that hold a stance that conflicts with s_i. For $i \in \{0..k\}$, let $O_i = \{p \in \Omega | \exists j \text{ s.t. } holds(p,s_j,T) \wedge conflict(s_i,s_j)\}$. The model with mutually exclusive stances can alternatively be expressed as:

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

Note that we are selecting with replacement, and it is possible for $p_1 = p_2$. 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:

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

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

$$P(conflicts|(s_i,s_j) = 1 \iff (i \neq j \wedge i \neq 0 \wedge j \neq 0)$$

This implies that $G_i \cap G_j = \emptyset$. Crucially, we set a lack of a stance to not 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[2]. Note that $|\Omega| = \Sigma_{i\in\{0..k\}}|G_i|$ and the probability of choosing any particular pair is $\frac{1}{|\Omega|^2}$. The denominator, $|\Omega|^2$, expands into the following expression:

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

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

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

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

and $P(nc|\Omega,T) = 1 - P(c|\Omega,T)$.

As before, we can trivially extend this definition to any non-empty sub-population $\omega \subseteq \Omega$ using $g_i = G_i \cap \omega$.

Trivially, $P(C|\omega,T)$ is maximal when $|g_0| = 0$ and $|g_1| = \ldots = |g_k| = \frac{|\omega|}{k}$, 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 $|g_0| = |g_1| = \ldots = |g_k| = \frac{|\omega|}{k-1}$.

Since the values of contention are $[0, \frac{k-1}{k}]$ rather than $[0,1]$, we normalize by the maximal contention (divide the contention score by $\frac{k-1}{k}$) and take the non-contention score as 1 minus the new score. This normalization brings both contention and non-contention to a full range of $[0,1]$, with a contention score of 1 signifying the highest possible contention regardless of the total number of stances.

3.2 Modeling Importance within Population

We now define a measure called *importance*. We loosely define importance as the level of impact that the issue brings to the world within a given population. In terms of importance of a topic T to population Ω, we interpret this as the number of people who think this topic is important to them. In other words, how many people are affected by T?

Let p be a person from some population and affected(T,p) be a binary function that returns whether p is affected by T. We let the probability that T is important to members of Ω be $P(I|\Omega,T)$. This is equivalent to the probability that T is important to the person p drawn from Ω.

$$P(I|\Omega,T) = P(p \text{ selected randomly from } \Omega \wedge affected(\text{p, T}))$$

Alternatively, we define Ω_T be the sub-population of Ω with those who are affected by T. $P(I|\Omega,T)$ can be computed by directly estimating $|\Omega_T|$.

$$P(I|\Omega,T) = \frac{|\Omega_T|}{|\Omega|}$$

Introducing importance into the controversy model allows us to penalize the controversy score for the contentious yet less important topics. For example, a sub-population ω_D that is affected by "The Dress" case is unarguably smaller than a sub-population ω_B that is affected by the Brexit Referendum. It follows that if we set Ω to the general population, we can now directly compare the importance of two topics in a general sense.

The *affected*(p) relationship is only defined conceptually here because its interpretation will vary with each dataset. In our experiments, we describe how to directly estimate the size of $|\Omega_T|$ from various datasets.

4 MODEL VALIDATION

We apply our model to the various data sources. To apply our theoretical model, we instantiate the model to algorithms that has been derived considering the characteristics of each dataset. We

examine three different data sources: polling data, Twitter, and Wikipedia. We show that our model has both explanatory power and predictive power.

4.1 Contention in Polling

In the Pew and Gallup datasets, we used the topline survey results as reported by each organization. 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.

4.1.1 US Scientists vs. General Population. Using one dataset acquired from Pew Research Center, a non-partisan fact tank in the US, we are able to examine attitudes towards a number of issues among two populations: US adults and US scientists. The opinions for US adults was gathered among a representative sample of 2,002 adults nationwide, while the opinions for scientists were gathered among a representative sample from the US membership of the American Association for the Advancement of Science (AAAS) (Table 2)

As seen in Figure 2, for some topics, such as offshore drilling, hydraulic fracturing (fracking), and biofuel, contention was similar between US 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 US adults. This result precisely matches earlier work's intuitive notion of politically, but not scientifically, controversial topics [27]. 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 dataset (98% of AAAS members surveyed said that "humans and other living things have evolved over time", whereas 31% of the US adults said that they have "existed in their present form since beginning of time").

4.1.2 Per-state distribution of Contention in the U.S. We obtained a dataset from the iSideWith.com website, a nonpartisan Voting Advice Application [5] 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 dataset by request from the website owners, which includes nation-wide and per-state opinions over 52 topics. Each topic is posed as a question with two main options for answers, usually a simple "yes" or "no". Additionally, the dataset includes the average importance of the issue (both nation-wide and per-state) rated by the users.

Using the iSideWith dataset, we measure 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?" ($P(c|US, t)$ = 0.26), and "Should every person purchasing a gun be required to pass a criminal and public safety background check?" ($P(c|US, t)$ = 0.39). Several topics had over 0.99 contention nation-wide, such as "Should the US 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

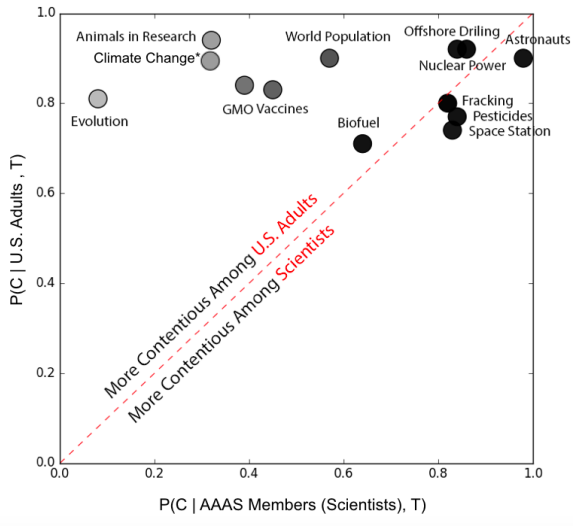

Figure 2: Contention in the scientific community vs. the general population for several controversial topics. The x=y line represents equal contention among both populations, with dots shaded according to their distance from the line. Note that the Climate Change question had 3 explicit stances and all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

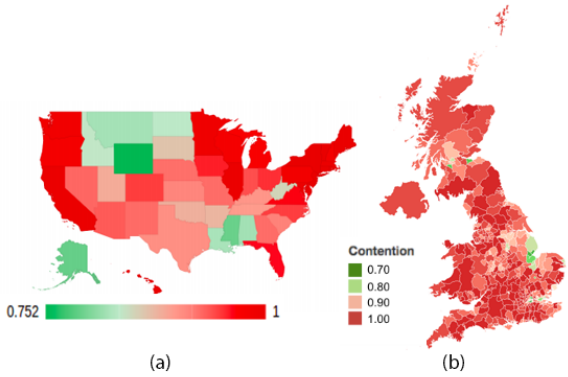

Figure 3: (a) Per-state contention for "Do you support increased gun control?". (b) Contention by voting district in the UK (The Electoral Comission 2016) Interactive maps for all iSideWith issues are available at `http://ciir.cs.umass.edu/irdemo/contention`

per-state contention for one such topic in Figure 3, 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`.

4.2 Controversy in Twitter

Social media allows people to quickly respond to a topic, compared to surveys or other types of media. We turn to Twitter and observe how controversy changes over time on three well-known

Table 2: Datasets containing explicit stances

Dataset	Type	# Issues	Population(s)	Years	# People	Source
Pew Adults	Statistically Calibrated Phone Survey	13	US adults	2014	2.0K	[20, 21]
Pew AAAS	Statistically Calibrated Online Survey	13	US scientists	2014	3.7K	[20, 21]
iSideWith	Informal Online Polling	52	US people	2014	varies (M)	By request
Dress Buzzfeed	Informal Online Polling	1	Online readers	2015-2016	3.5M	[14]

contentious topics: "The Dress", the Brexit Referendum, and 2016 US Presidential Election.

To apply our model to Twitter, we create an instance of our model that automatically identifies a population for whom the topic is relevant, and identify stances of people on the topic to measure contention. Note that we need to identify the population of interest first to estimate $|g_0|$, the group that has an interest in the topic but that does not take any stance.

4.2.1 Measuring Contention in a Twitter Population. We instantiate our model to compute the level of controversy for a given topic on Twitter. Instead of considering the entire Twitter population, we first automatically identify the sub-population interested in the topic T since the vast majority of people hold no stance even for the most controversial topics. Measuring contention within the relevant sub-population gives a more meaningful result on Twitter.

We start with a single hashtag "seeding" the topic, and the algorithm consists of three steps to identify the relevant sub-population and compute contention: (1) query hashtag expansion, (2) finding a population of interest, and (3) estimating the size of stance groups.

Query hashtag expansion: We start with the hashtag h of interest. Our goal here is to expand h to a set of k hashtags that are topically related to h, namely H_T

To do that, let $\mathcal{T}$ be a collection of tweets (e.g., the tweets collected for some day) and let $\mathcal{T}(h)$ be the subset of those tweets that contain the hashtag h. For any hashtag $h' \neq h$ that occurs in $\mathcal{T}(h)$, we calculate a TFIDF score as follows: TF is the number of times that h' occurs in $\mathcal{T}(h)$ and IDF is an inverse fraction of the frequency of h' in $\mathcal{T}$.

Note we aim to the topic T as the top k relevant hashtags H_T ranked by TFIDF in $\mathcal{T}$. However, one concern of that approach is that the hashtags in H_T is likely to vary greatly depending on which of them is chosen as a seed. To mitigate that risk, we create H_q for each hashtag $q \in H_T$. We then select the k hashtags that appear most often across all sets H_q. We call the resultant list T, and create a dataset $\mathcal{T}(T)$, which is a collection of tweets that contain any hashtag in T.

Identifying the population of interests to T: From $\mathcal{T}(T)$, we extract every user id who tweeted, is mentioned, or is retweeted. We consider this set of users as the population that shows interests in T. We call this sub-population ω_T as the people who are affected by T. ω_T is the population where the importance of T is maximized to 1 because by construction, it is the group of people who showed interests in T by explicitly discussing it on Twitter. Table 3 contains the size of $\mathcal{T}(T)$ and the identified population that shows interests.

Stance detection in the sub-population Automatic stance detection is an open problem [7, 12], so we use a simple and straightforward manual hashtag-based stance detection heuristic. We manually identified hashtags that explicitly indicate a stance. The full list of hashtags used is available at `http://ciir.cs.umass.edu/irdemo/contention`.

Table 3: Twitter dataset with implicit stances

Topic	# Tweets	# Users	Dates
The Dress	359K	361K	Feb. 26-Mar. 3, 2015
Brexit Referendum	14.8M	12.4M	May. 1-Jul. 24, 2016
US Elections	9.3M	6.2M	Sep. 20- Nov. 30, 2016
Total	24.4M	18.9M	

This high-precision, low-recall process will omit some tweets that do not use precisely the hashtags selected, but those that are selected are likely to be on the expected stance. 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 user ids in the tweets that contain any hashtag from each stance. As an estimate 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 users collected $-|G_1| - |G_2|$.

4.2.2 Controversy Trends on Twitter. We compute the final level of controversy by multiplying the contention computed on the identified population by the importance of that topic within the entire population that tweeted the same day. Figure 4 shows the controversy among all daily tweets by date for "The Dress," Brexit, and the 2016 US election. In all three plots, it shows marked peaks of contention around notable event times. For example, in the case of the US election, small peaks appear 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 relevant.

We compare $P(c|G_1 \cup G_2, T)$ from Twitter across a series of dates, along with the same probability calculated from external sources form comparison: the Buzzfeed poll on "The Dress" ($P(c|G_1 \cup G_2, T)$ = 0.88) [14], voting results on Brexit ($P(c|G_1 \cup G_2, T)$ = 1.00) [25], and the popular vote in the US Elections measured for the two main candidates ($P(c|G_1 \cup G_2, T)$ = 0.89). Additionally, Figure 3(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 ($P(c|Gibraltar, Brexit) = 0.16$). The extremely low contention makes sense: Gibraltar is geographically located inside Europe, and 95.9% of its voters voted "remain".

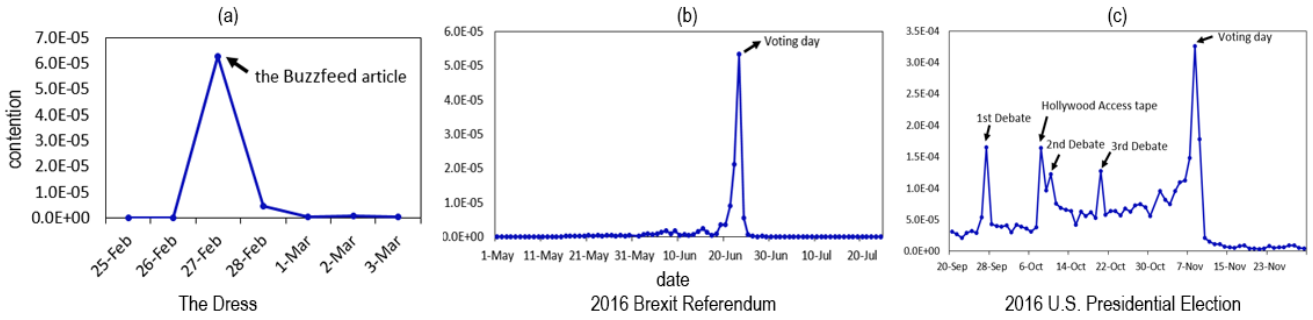

Figure 4: Controversy among all daily tweets by date for The Dress (left), Brexit (center) and 2016 US Elections (right), reported among all Gardenhose tweets that day (top) or only among those with an explicit stance (bottom). Notable peaks are annotated with associated events around that time. All dates are in UTC.

4.3 Controversy in Wikipedia

We now apply our model to the context of Wikipedia by measuring controversy among Wikipedia editor population.

4.3.1 Contention from Wikipedia Editor-population. Rather than estimating stances, our challenge now becomes to provide an estimate for the *conflicts* function directly between pairs of editors. Several past researchers have noted the centrality of Wikipedia reverts to the study of controversies [4, 23, 29]. Yasseri et al. in particular established reverts as a central mechanism for detecting controversy-related disagreement in Wikipedia [29].

Let $\mathfrak{W} = \{D_1, D_2, D_3 ... D_n\}$ be the collection of articles in Wikipedia, and let p be the person (i.e., the editor) that instituted any change to a document (such as insertions, deletions, and substitutions).

Let $E_D = \{e_1, e_2, ...e_k\}$ be the set of k edits applied to the document D. 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.

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 naive. We can conceptually distinguish between two types of reverts: those reverting vandalism and those reflecting opposing stances. To address this issue, Sumi et al., devised a reputation factor per editor, which grows proportionally with the number of edits the user contributes to this specific article [23].

The likelihood of an editor being a vandal is independent of all other editors. Adopting a probabilistic approach, we can reformulate 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

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

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 state 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 that 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 mutually reverted each other. We can then calculate contention as follows:

$$P(c|\Omega, D) = \frac{\sum_{p_1, p_2 \in \omega_D} P(conflicts(p_1, p_2))}{|\Omega|^2} = \frac{1}{|\Omega|^2} * \sum_{(p_i, p_j) \in MR_D} \frac{N_{p_i,D}}{N_{max}^D + 1} * \frac{N_{p_j,D}}{N_{max}^D + 1}$$

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.

Table 4: AUC measure reported on ranking controversial articles in Wikipedia by four scores.

	M [23]	C	MI	**CI**
AUC	0.649	0.649	0.630	**0.660**

4.3.2 Importance. We assume that an editor p who makes a change to the document is affected by the corresponding topic. Hence, we estimate $|\Omega_T|$ to be the size of the editors who have been involved with any change of the document.

$$P(I|T, \Omega_{\mathfrak{W}}) = \frac{|\omega_D|}{|\Omega_{\mathfrak{W}}|} \quad (1)$$

4.3.3 Ranking Controversial Articles in Wikipedia. We compare the contention derived from our model ("C") and controversy ("CI") scores, which is a version of C score multiplied by importance "I", against the state-of-the-art heuristic "M" score [23]. To verify the effectiveness of our new controversy score in predicting controversial Wikipedia articles, we rank all Wikipedia articles by four controversy-indicative scores: M, C, MI, and CI. To observe the effect of the importance score, we also devise an "MI" score, which is the M score weighted by its topic importance in Wikipedia. We used the truth data judgment for controversial Wikipedia articles from "the list of controversial issues" page [3] in Wikipedia as well as a previously collected annotated dataset [10]. Our judgment data contain 1,551 controversial articles. All articles that are not in the list are considered to be non-controversial. We then compute Area Under Curve (AUC) measure on the generated list.

Table 4 shows the AUC measure reported on ranking controversial articles by the four scores. While M and our C scores are comparable, CI score produced a better ranking than any of the measure. This result demonstrates that our model, when applied to Wikipedia, shows a competitive predictive power in classifying controversial articles in Wikipedia.

5 CONCLUSIONS

In this paper, we propose a theoretical model for controversy with respect to population. We argue that controversy is a multi-dimensional quantity that should only be understood in a given population and propose a model with two minimal dimensions: contention and importance. Contention mathematically quantifies the notion of "the proportion of people disagreeing on this topic" in a population-dependent fashion. On the other hand, importance measures how many people are affected by the given topic in the population. We validate our theoretical model on a wide variety of datasets from both offline and online sources, ranging from large informal online polls and Twitter data to statistically calibrated phone surveys and Wikipedia. Our experimental results show that our model has an explanatory power as well as a predictive power for the observed phenomenon.

6 ACKNOWLEDGEMENTS

The authors thank Marc-Allen Cartright, Jeff Dalton, Shay Hummel, Kiran Garimella, Seth Goldman, Justin Gross, Daniel Mishori, Brendan O'Connor, and Alena Vasilyeva for fruitful conversations. Special thanks to Taylor Peck and Nick Boutelier for providing us the iSideWith data set. This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-1217281. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

[3] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

REFERENCES

[1] L. Amendola, V. Marra, and M. Quartin. 2015. The evolving perception of controversial movies. *Palgrave Communications* 1 (2015).
[2] R. Awadallah, M. Ramanath, and G. Weikum. 2012. Harmony and Dissonance : Organizing the People's Voices on Political Controversies. *New York* (feb 2012), 523–532.
[3] E. Borra, A. Kaltenbrunner, M. Mauri, U. Amsterdam, E. Weltevrede, D. Laniado, R. Rogers, P. Ciuccarelli, and G. Magni. 2015. Societal Controversies in Wikipedia Articles. *Proceedings CHI 2015* (2015), 3–6.
[4] U. Brandes, P. Kenis, J. Lerner, and D. Raaij. 2009. Network analysis of collaboration structure in Wikipedia. WWW.
[5] L. Cedroni. 2010. Voting Advice Applications in Europe: A Comparison. *Voting Advice Applications in Europe: The State of Art* (2010), 247–258.
[6] Z. Chen and J. Berger. 2013. When, Why, and How Controversy Causes Conversation. *Journal of Consumer Research* 40, 3 (2013), 580–593.
[7] M. Coletto, C. Lucchese, S. Orlando, and R. Perego. 2016. Polarized user and topic tracking in twitter. In *SIGIR*. ACM, 945–948.
[8] P.A. Cramer. 2011. *Controversy as News Discourse*. Springer Netherlands.
[9] Marcelo Dascal. 1995. Epistemology, Controversies, and Pragmatics. *Isegoría* 12, 8-43 (1995).
[10] S. Dori-Hacohen and J. Allan. 2013. Detecting controversy on the web. In *CIKM*.
[11] F.H.V. Eemeren and B. Garssen. 2008. *Controversy and confrontation: Relating controversy analysis with argumentation theory*. Vol. 6. John Benjamins Publishing.
[12] K. Garimella, Aristides Morales, G.D.F., and M. Mathioudakis. 2016. Quantifying controversy in social media. *WSDM* (2016).
[13] D. J. Helfand. 2016. *A Survival Guide to the Misinformation Age: Scientific Habits of Mind*. Columbia University Press.
[14] C. Holderness. 2015. What Colors Are This Dress? (2015). https://www.buzzfeed.com/catesish/help-am-i-going-insane -its-definitely-blue.
[15] M. Jang and J. Allan. 2016. Improving Automated Controversy Detection on the Web. In *SIGIR*.
[16] M. Jang, J. Foley, S. Dori-Hacohen, and J. Allan. 2016. Probabilistic Approaches to Controversy Detection. In *CIKM*.
[17] A. Kittur, B. Suh, B.A. Pendleton, and E. H h. Chi. 2007. He says, she says: conflict and coordination in Wikipedia. In *CHI*.
[18] M. Klenner, M. Amsler, and N. Hollenstein. 2014. Verb Polarity Frames: a New Resource and its Application in Target-specific Polarity Classification. In *KONVENS*.
[19] A.I. Leshner. 2015. Bridging the opinion gap. *Science* 347, 6221 (2015), 459.
[20] Pew Research Center. 2015. *An Elaboration of AAAS Scientists' Views*. Technical Report.
[21] Pew Research Center. 2015. *Public and Scientists' Views on Science and Society*. Technical Report.
[22] H. S. Rad and D. Barbosa. 2012. Identifying controversial articles in Wikipedia: A comparative study. In *WikiSym '12*. ACM.
[23] R. Sumi, T. Yasseri, A. Rung, A. Kornai, and J. Kertész. 2011. Edit Wars in Wikipedia. In *2011 IEEE Third Int'l Conference on Social Computing*.
[24] Mihály Szívós. 2005. Temporality, reification and subjectivity. *Controversies and Subjectivity* 1 (2005), 201.
[25] The Electoral Comission. 2016. EU referendum results. (2016).
[26] B. Timmermans, L. Aroyo, T. Kuhn, K. Beelen, E. Kanoulas, B. van de Velde, and G. van Eerten. 2017. ControCurator: Understanding Controversy Using Collective Intelligence. *Collective Intelligence* (2017).
[27] A. M Wilson and G. E Likens. 2015. Content volatility of scientific topics in Wikipedia: A cautionary tale. *PLoS ONE* 10, 8 (2015), 10–14.
[28] T. Yasseri, A. Spoerri, M. Graham, and J. Kertész. 2014. The most controversial topics in Wikipedia: A multilingual and geographical analysis. In *Global Wikipedia: International and cross-cultural issues in collaboration*. 178.
[29] T. Yasseri, R. Sumi, A. Rung, A. Kornai, and J. Kertész. 2012. Dynamics of conflicts in Wikipedia. *PloS one* 7, 6 (Jan. 2012), e38869.