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. As 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 systematic understanding of how controversy should be defined and measured. In this paper, we propose a multi-dimensional model of controversy as a systematic way to understand it. Specifically, we introduce a model with two minimal dimensions, "contention" and "importance". Our model departs from other work by viewing controversy as trait rooted in population. It suggests that controversy should be separately observed in a given population, rather than 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 predictive power for tasks such as identifying controversial Wikipedia articles.

CCS CONCEPTS

•**Computer systems organization** →**Embedded systems;** *Redundancy;* Robotics; •**Networks** →Network reliability;

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference format:
Myungha Jang, Shiri Dori-Hacohen and James Allan. 1997. Modeling Controversy within Populations. In *Proceedings of ACM Woodstock conference, El Paso, Texas USA, July 1997 (WOODSTOCK'97),* 9 pages.
DOI: 10.475/123_4

1 INTRODUCTION

Social network tools such as Twitter, Facebook, discussion forums, and comments on news articles are increasingly the place where controversial arguments are being held. Technological tools have become critical in shaping these discussions by influencing which users see which data, through algorithmic curation and filtering. 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, and has made significant strides in analyzing and detecting controversy (cf. [3, 11]). 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 select topics of broad interest and 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 [18], with many scientists explicitly fighting these trends by insisting "there is no controversy" [12] (referring to *scientific* controversy). 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]?

We address these issues by proposing a theoretical model that defines controversy in terms of population and as a combination of (at least) "contention" and "importance". The model thus 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 "is climate change controversial to $\{$*a particular group*$\}$?".

Our framework departs from most existing work about controversy in several major ways. First, we define controversy 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 any number of stances rather than just two opinions. Third, our model allows that some items may be less controversial because they are contentious but not important (or 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,

WOODSTOCK'97, El Paso, Texas USA

DOI: 10.475/123_4

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

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 [19]. Additionally, we look at Twitter coverage for 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. Lastly, we apply our model to Wikipedia. We show that our model also has a predictive power in classifying controversial Wikipedia articles with the metric derived from our model.

2 RELATED WORK

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 [2, 3, 21]. With few exceptions [1, 14], prior work did not model controversy formally. Even when it did, the meaning of controversy was not modeled, 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.[15, 27]). Past research shows that achieving inter-annotator agreement on the "controversy" label is challenging [10, 16].

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. [23, 25]), or else works towards a more qualitative analysis of the overall patterns across controversies [9], with one notable exception [8]. In philosophy, Leibniz offered a simple definition of controversy: a controversy is a question over which contrary opinions are held [17], which Dascal notes as "clearly insufficient" [9]. Dascal offers a theory of controversies which distinguishes between types of polemic discourse [9]. 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 frivolous topics such as the colors of a dress.

We depart from past research by modeling controversy as a multi-dimensional quantity, of which "contention" and "importance" are minimal dimensions and which accounts for such differences. Similarly, Timmermans et al., also identified five aspects of controversy in news articles, such as time persistence, emotion, multitude of actors, polarity and opennesss. However, their approach is mainly targeted for news articles and has less focused on actual modeling

3 MODELING POPULATION-BASED CONTROVERSY

"Controversy" is a complicated social phenomenon. As it is difficult to formally and systematically define what the controversy is, there has been little efforts to formulate models that quantify the level of controversy for the given topic.

As a motivational example, consider two controversies of "The Dress" and Brexit referendum. "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 [13]. 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.

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 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 predict effects on diplomatic relationships and the world economy for years to come. 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. Its impact on the world was likely negligible.

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 is insufficient to explain such arguably frivolous controversies as . 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. Using the same notation as above, we hypothesize that these are minimal dimensions of controversy.

This framework is demonstrated schematically in Figure 1, overlaying actual results including importance reported in the iSideWith data set (see Table 1). The first dimension is "contention" which we defined 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 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 affordable care act have high contention and high importance, hence more controversial. Issues such as whether the government should provide incentives for trucks

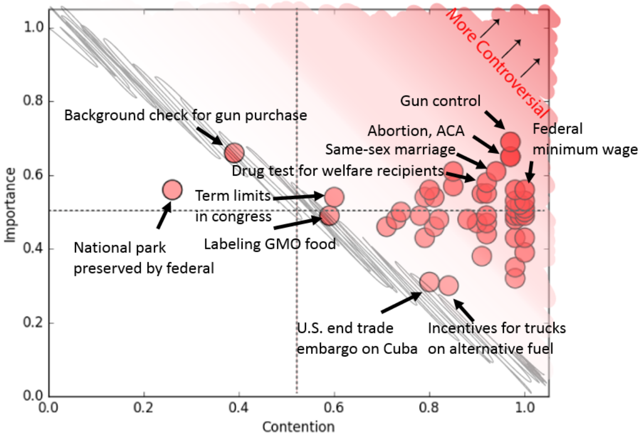

Figure 1: iSideWith topics plotted reconceptualizing controversy as composed of at least two dimensions, contention and importance. Sample topics are given in each quadrant of {low,high} importance and contention.

to run on alternative fuels is highly contentious but is rated by users as low importance.

Finally, we model the probability of controversy with given T and within a given population Ω as $P(Controversy|\theta)$, where $\theta = \{T, \Omega\}$. Our model hypothesizes that the probability of controversy given T and Ω is the joint probability of two aspect, contention (C) and importance (I).

$$P(Controversy|\theta) = P(\mathbf{C}ontention, \mathbf{I}mportance|\theta)$$

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

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

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

We now discuss the modeling of $P(C|\theta)$ and $P(I|\theta)$ and how to estimate them from the real data.

3.1 Modeling Contention from Population

We mathematically formulate 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 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 Ω. We define c to denote the level of contention, which we 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 $P(nc|\Omega, T)$ 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 $P(conflict|s_i, s_j)$ be a probability that s_i and s_j are in a complete conflict. This probability measures the severity of conflict that two stances are.

For example, if two stances s_i and s_j are in a complete conflict, which means that two stances are mutually exclusive. We assume that a person that takes s_i does not s_j on T when s_i and s_j are mutually exclusive. The examples of the mutually exclusive stances are pro-Hillary Clinton and pro-Donald Trump for 2016 U.S. presidential election, or two stances of "abortion should not be legalized" and "abortion should be legalized" for abortion. The $P(conflict|s_i, s_j)$ return the maximal value of 1 when two stances are in a mutually exclusively conflict.

However, not all stances are mutually exclusive. 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 stance distance between the third stance and the other two stance should be lower than 1.0. Note that a person can hold multiple stances simultaneously as long as any of the two stances are mutually exclusive. However, no stance can be jointly held with s_0. We set $P(conflict|s_i, s_i) = 0$ and $P(conflict|s_0, s_i) = 0$

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, $\Omega = \bigcup_i G_i$. We let $P(conflict|G_i, G_j)$ be a probability that two groups of G_i and G_j are in a complete conflict, similarly as it was defined on the 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:

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

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. The probability of contention with mutually exclusive stances is a special case in our model where $P(conflict|s_i, s_j)$ is 1. Most of controversial topics have at least two exclusive mutually stances to bisect the community. In this section, we focus more on the case of mutually exclusive stances and describe the model can be instantiated. The model described here can be easily generalized by adjusting the value of $P(conflict|s_i, s_j)$.

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

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

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:

$$conflicts(s_i, s_j) \iff (i \neq j \wedge i \neq 0 \wedge j \neq 0)$$

This implies that $G_i \cap G_j = \emptyset$. Crucially, we set a lack of 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[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:

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

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,

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

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 $g_i = G_i \cap \omega$. 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, $|\hat{S}| \leq 1 \implies P(c|\Omega,T) = 0$.

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 $[0, \frac{k-1}{k}]$ 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 $[0,1]$ each, 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 formulate a measure called "importance". We loosely define "importance" as the level of impact that the issue brings to the world within the population. In terms of importance of T to 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 as $P(I|\Omega,T)$. This is equivalent to the probability that T is important to the person p drawn from Ω.

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

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 $|\Omega_T|$.

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

How to define the function affected(p) or estimating the size of $|\Omega_T|$ can vary on the dataset. We discuss how we estimate the size of $|\Omega_T|$ from different datasets, such as Twitter and Wikiepdia.

4 MODEL VALIDATION

We apply our model to the various data sources. To apply our theoretical model, we instantiate the model to algorithms that reflects different characteristics of each dataset. We examine three different data sources, polling data, Twitter, and Wikipedia. We validate our model by showing that it has both explanatory power and predictive power.

4.1 Contention in Polling

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.

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

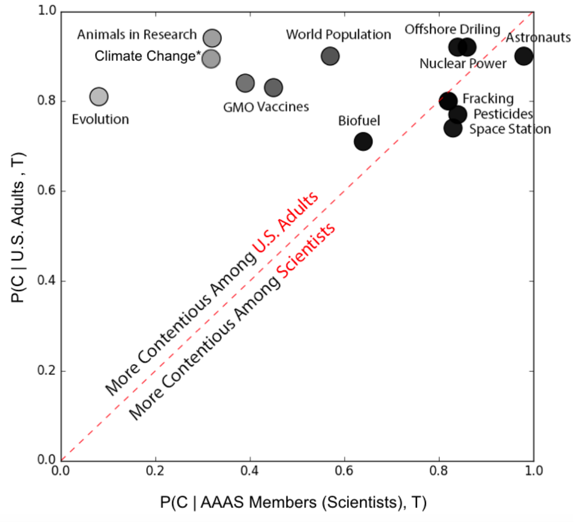

Figure 2: Contention in the scientific community vs. general population for several controversial topics. The x=y line represents equal contention among both populations, with dots shaded according to their distance from the line. Note that the Climate Change question had 3 explicit stances, all other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

4.1.1 U.S. Scientists vs. General Population.

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. The opinions for U.S. 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 U.S. membership of the American Association for the Advancement of Science (AAAS) (Table 1)

As seen in Figure 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 [26]. 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 surveyed 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").

4.1.2 Per-state distribution of Contention in the U.S.

We obtained a data set 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 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.

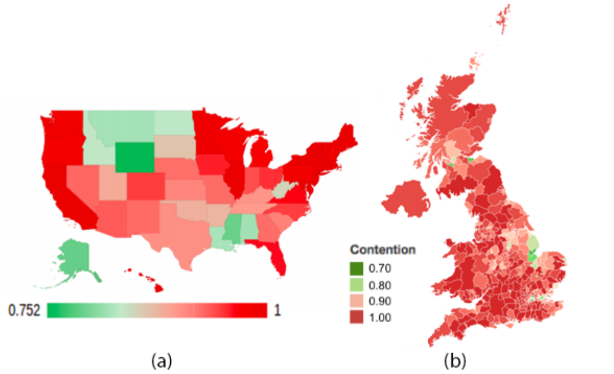

Figure 3: (a) Per-state contention for "Do you support increased gun control?". (b) Contention by voting district in the UK (The Electoral Comission 2016) Interactive maps for all iSideWith issues are available at [[redacted for review]]

Using the iSideWith data set, we measured 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 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 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 [redacted for review].

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 contentious topics: The Dress, Brexit Referendum, and 2016 U.S. Presidential Election.

4.2.1 Measuring Contention in a Twitter Population.

We instantiate our model to compute the level of controversy of a given topic in Twitter. We start with a single hashtag "seeding" the topic and the algorithm consists of three steps: (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. We expand h to a set of k hashtags that are topically related to h, which we call topic T_h.

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

Table 1: Data sets containing explicit stances

Dataset	Type	# Issues	Population(s)	Years	# People	Source
Pew Adults	Statistically Calibrated Phone Survey	13	US adults	2014	2.0K	[19, 20]
Pew AAAS	Statistically Calibrated Online Survey	13	US scientists	2014	3.7K	[19, 20]
iSideWith	Informal Online Polling	52	US people	2014	varies (M)	By request
Dress Buzzfeed	Informal Online Polling	1	Online readers	2015-2016	3.5M	[13]

Table 2: Example hashtags used to identify two stance groups on The Dress, Brexit and the U.S. Elections. Full list at [[redacted for demo]].

Topic	Stances	Example Hashtags	# of hashtags
The Dress	Blue and Black	#blackandblue, #notwhiteandgold, #blackandbluedress,#青と黒,#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

Table 3: Twitter Data set 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
U.S. Elections	9.3M	6.2M	Sep. 20- Nov. 30, 2016
Total	24.4M	18.9M	

$\mathcal{T}(h)$, we calculate a TFIDF score as follows: TF is the number of times that h' occurs in $\mathcal{T}(h)$ and IDF is the inverse of the number of hashtags h'' such that h' is contained in $\mathcal{T}(h'')$.

We let set of the top k hashtags ranked by TFIDF be T_h. One concern of that approach is that the hashtags in T_h is likely to vary greatly depending on which of them is chosen as a seed. To mitigate that risk, we create T_q for each hashtag $q \in T_h$. We then select the k hashtags that appear most often across all sets T_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, or 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 as 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 a open problem [7, 11], so we use a simple and straightforward manual hashtag-based stance detection heuristic. We manually identified hashtags that explicitly indicate a stance. As examples, Table 2 shows the hashtags we used to identify two mutually exclusive stances in three contentious topics. 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 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 $-|G_1| - |G_2|$.

4.2.2 Controversy Trends on Twitter. We compute the final level of controversy by multiplying the contention computed 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 2016 U.S. election. In all three plots, it shows marked peaks of contention around notable event times. For example, in the U.S. Elections case, 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 salient.

We compare $P(c|G_1 \cup G_2, T)$ from Twitter across a series of dates, with that calculated from external sources: the Buzzfeed poll on The Dress ($P(c|G_1 \cup G_2, T) = 0.88$) [13], voting results on Brexit ($P(c|G_1 \cup G_2, T) = 1.00$) [24], and the popular vote in the U.S. 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".

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.

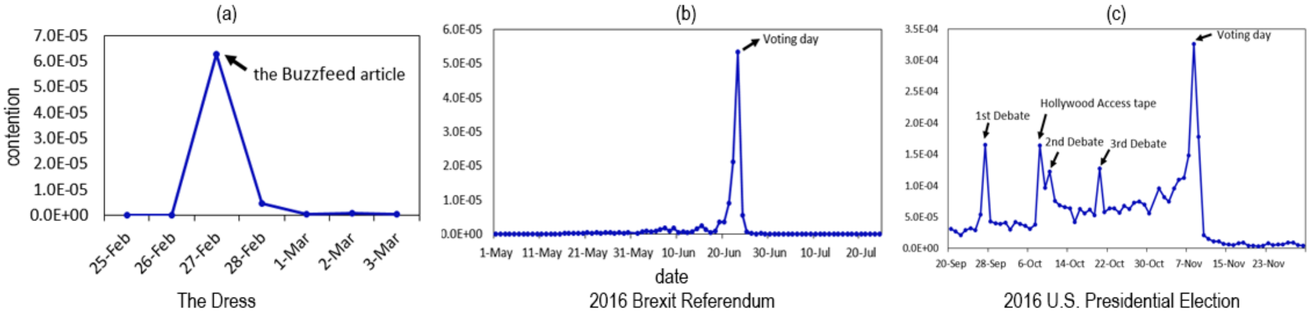

Figure 4: Controversy among all daily tweets by date for The Dress (left), Brexit (center) and 2016 U.S. Elections (right), reported among all Gardenhose tweets that day (top) or only among those with an explicit stance (bottom). Notable peaks are annotated with associated events around that time. All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

Several past researchers have noted the centrality of Wikipedia reverts to the study of controversies [4, 22, 28]. Yasseri et al. in particular established reverts as a central mechanism for detecting controversy-related disagreement in Wikipedia [28].

Let $\mathfrak{W} = \{D\}$ be the collection of articles in Wikipedia. and p is the person (editor) that instituted any change to a document (such as insertions, deletions, and subsitutions).

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 naïve. We can conceptually distinguish between two types of reverts: those reverting vandalism and those reflecting opposing stances. 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 [22, 28]. The first heuristic they proposed to focus exclusively on **mutual reverts**, i.e. cases in which both editors have reverted each other. Let $conflicts_{mr}(p_1, p_2) \equiv reverts(p_1, p_2) \wedge reverts(p_2, p_1)$.

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 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 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:

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

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:

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

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

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

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

	M [22]	C	MI	**CI**
AUC	0.649	0.649	0.630	**0.660**

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 's editors.

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 $|\Omega_T|$ be the size of the editors who have been involved with any change of the document.

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

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 [22]. We rank Wikipedia articles by four controversy-indicative scores, M, C, MI, and CI. To observe the effect of importance score, we also devise "MI" score, which is M score weighted by its topic importance in Wikipedia. We compute Area Under Curve (AUC) measure on the generated list. We used the truth data judgment for controversial Wikipedia articles from "the list of controversial issues" page [3] in Wikipedia as well as previously collected annotated dataset [10]. Our judgment data contain 1,551 controversial articles.

Table 4 shows the AUC measure reported on raking 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 results 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 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. This model allows us, for example, to formally answer the question in the title of our paper, "Is Climate Change Controversial?", differently depending on the population being observed: climate change is not contentious in the scientific community, yet is in the general U.S. public. We validate 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 Wikipedia. Our experimental results show that our model has an explanatory power for the observed phenomenon as well as predictive power.

6 ACKNOWLEDGEMENTS

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.

REFERENCES

[1] Luca Amendola, Valerio Marra, and Miguel Quartin. 2015. The evolving perception of controversial movies. *Palgrave Communications* 1 (2015).

[2] Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. 2012. Harmony and Dissonance : Organizing the People's Voices on Political Controversies. *New York* (feb 2012), 523–532.

[3] Erik Borra, Andreas Kaltenbrunner, Michele Mauri, Uva Amsterdam, Esther Weltevrede, David Laniado, Richard Rogers, Paolo Ciuccarelli, and Giovanni Magni. 2015. Societal Controversies in Wikipedia Articles. *Proceedings CHI 2015* (2015), 3–6.

[4] Ulrik Brandes, Patrick Kenis, Jurgen Lerner, and Denise van Raaij. 2009. Network analysis of collaboration structure in Wikipedia. WWW.

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

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

[7] Mauro Coletto, Claudio Lucchese, Salvatore Orlando, and Raffaele 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] Shiri Dori-Hacohen and James Allan. 2013. Detecting controversy on the web. In *CIKM*.

[11] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying controversy in social media. *WSDM* (2016).

[12] David J Helfand. 2016. *A Survival Guide to the Misinformation Age: Scientific Habits of Mind*. Columbia University Press.

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

[14] Myung-ha Jang, John Foley, Shiri Dori-Hacohen, and James Allan. 2016. Probabilistic Approaches to Controversy Detection. In *CIKM*.

[15] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed Huai hsin Chi. 2007. He says, she says: conflict and coordination in Wikipedia. In *CHI*.

[16] Manfred Klenner, Michael Amsler, and Nora Hollenstein. 2014. Verb Polarity Frames: a New Resource and its Application in Target-specific Polarity Classification. In *KONVENS*.

[17] G W Leibniz. 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 (1982), 1253.

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

[19] Pew Research Center. 2015. *An Elaboration of AAAS Scientists' Views*. Technical Report. http://www.pewinternet.org/2015/07/23/an-elaboration-of-aaas-scientists-views/

[20] Pew Research Center. 2015. *Public and Scientists' Views on Science and Society*. Technical Report. http://www.pewinternet.org/2015/01/29/public-and-scientists-views-on-science-and-society/

[21] Hoda Sepehri Rad and Denilson Barbosa. 2012. Identifying controversial articles in Wikipedia: A comparative study. In *WikiSym '12*. ACM.

[22] Róbert Sumi, Taha Yasseri, András Rung, András Kornai, and János Kertész. 2011. Edit Wars in Wikipedia. In *2011 IEEE Third Int'l Conference on Social Computing*.

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

[24] The Electoral Comission. 2016. EU referendum results. (2016). 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.

[25] Frans H Van Eemeren and Bart Garssen. 2008. *Controversy and confrontation: Relating controversy analysis with argumentation theory*. Vol. 6. John Benjamins Publishing.

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

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

[27] Taha Yasseri, Anselm Spoerri, Mark Graham, and János Kertész. 2014. The most controversial topics in Wikipedia: A multilingual and geographical analysis. In *Global Wikipedia: International and cross-cultural issues in collaboration*. 178.

[28] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. 2012. Dynamics of conflicts in Wikipedia. *PloS one* 7, 6 (Jan. 2012), e38869.