[Northeast Journal of Complex Systems (NEJCS)](https://orb.binghamton.edu/nejcs)

Volume 5
Number 1 *Special Issue: NERCCS 2023 Papers*

Article 2

June 2023

Characterizing Controversiality of Topics Utilizing Eccentricity of Opinions

Sriniwas Pandey
Binghamton University, spandey4@binghamton.edu

Hiroki Sayama
Binghamton University, SUNY, sayama@binghamton.edu

Characterizing Controversiality of Topics Utilizing Eccentricity of Opinions

Sriniwas Pandey[1*] and Hiroki Sayama[1,2]

[1] Center for Collective Dynamics of Complex Systems

Binghamton University, State University of New York, Binghamton, NY, USA

[2] School of Commerce, Waseda University, Tokyo, Japan

[*]spandey4@binghamton.edu

Abstract

Access to abundant biased information in echo chambers and social bubbles often intensifies opinions to the extremes. The extremization of opinions results in several topics becoming controversial. However, it is very difficult to measure the degree of controversiality of a topic objectively since the controversiality of any topic is subjective and perceived differently from different communities. The absence of an objective measure of controversiality has been a major hindrance in understanding the causes and effects of it. In this work we propose a method to quantify controversiality of a topic by utilizing eccentricity of opinions on that topic. The eccentricity of an opinion is the amount of strangeness of the opinion relative to other opinions in the social neighborhood. The collective eccentricity of all opinions for a topic works as an indicator of the controversiality of that topic and can be represented by any measure of central tendency. With the help of social network data, we also demonstrate that opinions on several issues related to our routine life show similar trends for diversity though they differ in their controversiality.

1 Introduction

In our society, the space for civilized discussions is increasingly shrinking due to the growth of strongly cohesive groups that have very little tolerance for disagreements. This has led to the labeling of several topics as controversial, and often, discussion on these topics is avoided in public forums to prevent disputes. However, understanding controversy and measuring it objectively is a complex task that has been a topic of interest to researchers for several years [1, 8, 9, 11, 15, 17, 21].

Most of the research on this topic can be classified into two broad categories. A common dimension in which several of these works focus is identifying controversy [9, 17, 3, 13] where they try to determine if any opinion or set of opinions is controversial or not. Another most popular dimension of these research works is quantifying or measuring the level of controversy [2, 26, 14, 16]. The research on controversy detection mainly utilizes data from online social networks and particularly Twitter [11, 21, 16, 20]. There are some works which analyze Reddit, Wikipedia, and other online news datasets as well [9, 17, 13]. The methods in both categories, i.e., controversy detection and quantification, are primarily network-based, content-based, or a combination of them [14]. Network-based methods primarily rely on the analysis of retweet and comment networks or networks of information citations. To gain a deeper understanding of these networks, cluster analysis and community detection algorithms are often employed. Additionally, content-based controversy detection and quantification algorithms are utilized to scrutinize the textual content of opinions. Sentiment analysis and opinion dynamics analysis are among the most prominent methods in this category [9, 22, 25]. Many studies focus on analyzing controversial political and religious topics [1, 20] , such as abortion, gun control, and religious freedom. These topics have received extensive attention and research [7, 10, 12, 23, 24] . Most existing works categorize topics as either controversial or non-controversial, and often assume that controversies are polarized, such as left vs right political ideology or Christianity vs Islam. Another common theme among these works is that controversiality is identified and measured at a single global level, assumed to be standard. However, it is essential to understand that controversiality is a subjective feature and must be evaluated in a social context or neighborhood. For instance, the topic of "eating beef" is not controversial at all in American society, but in some other countries like India, it is highly controversial. Moreover, we believe that binary classification into controversial or non-controversial categories is not always appropriate. Controversiality should be measured on a continuous scale. The assumption that controversiality is identified and measured at a single global level is flawed. Instead, it should be assessed in its social and cultural context.

Our proposed method seeks to overcome the limitations of existing works by introducing a more comprehensive approach to measuring controversiality. Instead of relying solely on the topic being discussed, our method incorporates the concept of social context by utilizing a user's social neighborhood to measure controversiality. This allows for a more flexible and personalized approach to measuring controversiality that accounts for the role of social context in shaping opinions and attitudes. Rather than classifying a topic as controversial or non-controversial, it quantifies the degree of controversiality in each topic discussion on a continuous scale. This allows for a more nuanced understanding of controversiality that can account for the varying degrees of controversy within a topic. Our algorithm is versatile, and opinions on any topic, subject, or keyword that has been discussed can be fed into it to measure the associated controversy. We also demonstrate that topics that are considered highly polarized are not necessarily homogeneous within an extreme community. Instead, they can be highly diverse. Overall, our approach provides a more comprehensive and nuanced perspective on controversiality that accounts for social context and the diverse nature of controversial topics.

2 Related Work

Zarate et al. in their work [26] utilize community detection on a retweet network to measure the controversy. Al-Ayyoub et al. [6] analyze Arabic tweets on trending topics from different domains to analyze the controversy. They use both network-based and content-based methods separately for this purpose. Wikipedia metadata has been used by Dori-Hacohen et al. [13] to automate the controversy detection task. Choi et al. [9] identify controversial issues and subtopics by measuring magnitude of sentiments for contradicting opinions. Popescu et al. [21] recognize the controversial events from Twitter using combinations of classification and regression models. In these models they use Twitter based features like linguistic, structural, sentiment etc. and External features such as News Buzz and Web News controversy. Guerra et al. [16] analyze the distribution of nodes at the boundary of communities to identify polarization. Jang et al. [17] apply probabilistic approaches using controversy language models to identify Wikipedia controversy. Mean and variance of sentiments is used to measure the contradictions of opinions by Tsytsarau et al. [25] Akoglu [2] in his work identifies and ranks the topics polarity solving a node classification problem using the bipartite networks. Al Amin et al. [3] in their work utilize a matrix factorization approach on a matrix of information source and social dependency network. Qiu et al. [22] create an emotional social network to partition nodes into three categories proponent, opponent and neutral using the simulated annealing algorithm. One of the recent works by Garimella et al. [14] use content and network structure of social networks to quantify controversy. By analyzing the partitions of the conversation graph and applying several controversy measuring methods they calculate controversy of the conversation graph for individual users. They present a detailed review of existing methods and literature about the controversiality detection.

3 Proposed Method

We start our method by cleaning the raw text opinions collected from the social network to remove contractions, digits, punctuation marks and special characters.

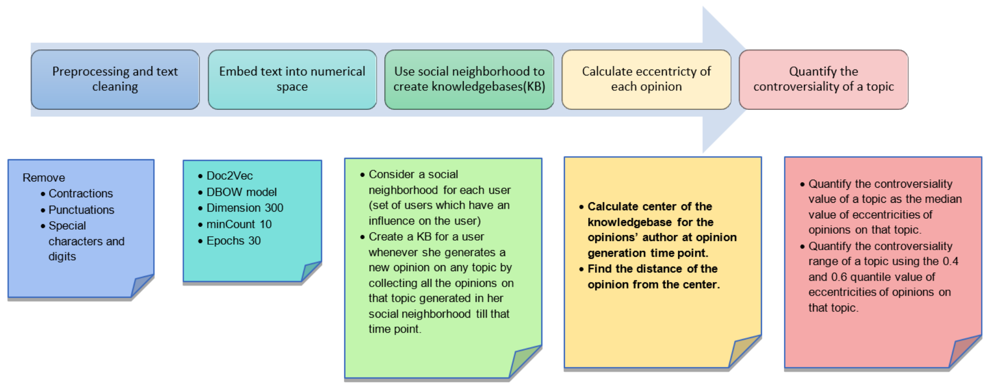

Figure 1: Schematic diagram representing the proposed method for controversiality quantification: Actions taken at each step are listed in the box below the step.

After the text preprocessing step, we apply the Doc2Vec algorithm [18] to embed text posts in a numerical vector space. Doc2Vec is a text embedding method developed upon the Word2Vec model [19], in which a neural network is trained using the given data to learn the document and word embeddings. In our work we have used the Distributed Bag of Words (DBOW) Doc2Vec model to convert cleaned text opinions into numerical vectors from the text. For training the model, words occurring less than 10 times in the dataset were ignored and training was done in 30 epochs. The model was trained to generate 300 dimensional vectors.

Once all the posts are represented in numerical vector form, the next step is to calculate the eccentricity of each post. We define eccentricity as the distance of an opinion from the center of all opinions in the social neighborhood. In other words, the eccentricity of an opinion is the measure of 'oddness' of that opinion in comparison to the norm of the neighborhood. Neighborhood of a user can be defined at different social contexts, for example a neighborhood can be the set of users a user often interacts with, or the set of all users she is following or being followed by or both, or all the users in a social group etc.

The eccentricity of an opinion is relative to the social context and calculated as follows: To calculate the eccentricity of an opinion $O^T_{(u,t)}$, which is an opinion on topic T posted at time t by user u, we first need to create a knowledge base $KB^T_u(t)$. The knowledge base of user u on topic T at time t, $KB^T_u(t)$, is the collection of all opinions on topic T in the social neighborhood of user u that were posted before time t. Eccentricity of the opinion $O^T_{(u,t)}$, denoted as $Ecc(O^T_{(u,t)})$, is the distance (d) of $O^T_{(u,t)}$ from the center of $KB^T_u(t)$.

$$KB^T_u\ (t) = \{O^T_{(v,k)} | v \in Neighborhood(u), k < t\ and\ (t-k) <\ 5\ days\} \quad (1)$$

$$Ecc(O^T_{(u,t)}) = d\,(O^T_{(u,t)},\ center(KB^T_u(t))) \tag{2}$$

Social context or neighborhood depends on the purpose and requirement of eccentricity analysis. Since the data in our analysis comes from a small community of specific ideological people, we have assumed all users in the network in each other's social neighborhood,$\forall_{(u \in V)}\ Neighborhood(u) = V$, where V is the set of all users in our dataset. The controversiality we get in this work will be in the context of Parler community. However, the proposed method does not dictate social context determination and updating the social context will measure the controversiality in that context. The distance measure we have used in our work is Euclidean distance and center of knowledge base is represented by the mean of opinions in $KB^T_u(t)$. Different social contexts, distance metrics and center representations can be used as per the requirement. According to the chosen settings for our work, equations 1 and 2 are simplified as follows. Knowledge base on any topic T will be same for every user in the network, hence $KB^T_u(t)$ is replaced by $KB^T(t)$.

$$KB^T(t) = \{O^T_{(v,k)} \mid v \in V\ and\ k < t\ and\ (t\ -\ k)\ <\ 5\ days\} \tag{3}$$

$$Ecc(O^T_{(u,t)}) = \|(O^T_{(u,t)},\ mean(KB^T(t)))\|^2 \tag{4}$$

Once we have eccentricity of all the opinions, we analyze the collective distribution of eccentricities for all the opinions on a topic. We observed that the eccentricity of all opinions for a topic characterizes the controversiality of that topic. We propose two measures derived from the collective distribution which can be used to measure the controversiality. The first measure we propose to quantify controversiality of a topic is the median value of eccentricity distribution of all the opinions on that topic.

$$Ecc^T\ =\ \{Ecc(O^T_{(u,t)}) \mid u \in V,\ \forall_t\} \tag{5}$$

$$Controversiality(T)\ =\ Median(Ecc^T) \tag{6}$$

This equation assigns a single number to a topic T which is the value of controversiality of the topic. Using median value to quantify controversiality is useful since this measure is not affected by the outliers and represents the central tendency of data. However, quantifying controversiality with a single number might be very strict in some scenarios, especially when the topics are very close in terms of controversiality. We will elaborate this further in the result section. To overcome this limitation, we propose another method to quantify controversiality where we assign each topic a controversiality range using the quantiles. We use the 0.4 - 0.6 quantile range of the eccentricity distribution to measure the controversiality.

$$Controversiality\,range\ (T) =\ (Q_{0.4}\ (Ecc^T)\ ,\ Q_{0.6}\ (Ecc^T)) \tag{7}$$

$$where\ Q_i\ (D)\ is\ the\ i^{th}\ quantile\ of\ data\ D.$$

Our approach of choosing the 0.4 – 0.6 quantile range as a controversiality measure is motivated by the following reasons. i) the proposed range is around the central tendency of data and remains unaffected by the outliers ii) it is more relaxed than a single number like median. However, other measures utilizing the eccentricity distribution e.g., mean eccentricity value of all opinions, mean value of top fifty percent eccentric opinions, or any other wider/narrower quantile range, can also be used depending upon the dataset and purpose of controversiality measurement. The schematic diagram of steps followed for our analysis is shown in Figure 1.

4 Dataset

For this study we analyzed Parler social network dataset with 183M posts and comments which were posted between August 2018 and November 2021 [4]. Parler is a microblogging and social networking site mostly popular among the alt-right, Trump supporters and believers of conspiracy theories [5]. We believe that if we can detect a connection between controversy and eccentricity of opinion in this dataset, where opinions are already on an extreme, it will be implied that similar patterns exist in other mainstream social networks and datasets. From the Parler dataset, we selected six topics for our analysis: i) Shirt, ii) Chocolate, iii) Vegetarian, iv) Climate change, v) Vaccine, and vi) Gun. We chose these topics because we know they have varied levels of controversiality: Shirt and Chocolate were included as a baseline with minimum controversy, while Gun and Climate change as extremely controversial topics. For each topic, posts were filtered using a simple keyword search by searching for a single keyword (similar as topic name) related to the topic. The distribution of the numbers of posts and authors for each topic after initial preprocessing is shown in Table 1. The topics are diverse not only in content but also in terms of total numbers of posts and authors.

Topic	**# of gabs(posts)**	**# of authors**
Chocolate	7,931	5,661
Climate change	21,221	11,510
Gun	186,851	70,163
Shirt	96,981	51,174
Vaccine	468,085	142,561
Vegetarian	1,721	1,301

Table 1: Number of posts and their authors in the preprocessed dataset: Topics are diverse in terms of number of opinions and authors. Vegetarianism is the least discussed topic while Vaccine is the most popular one.

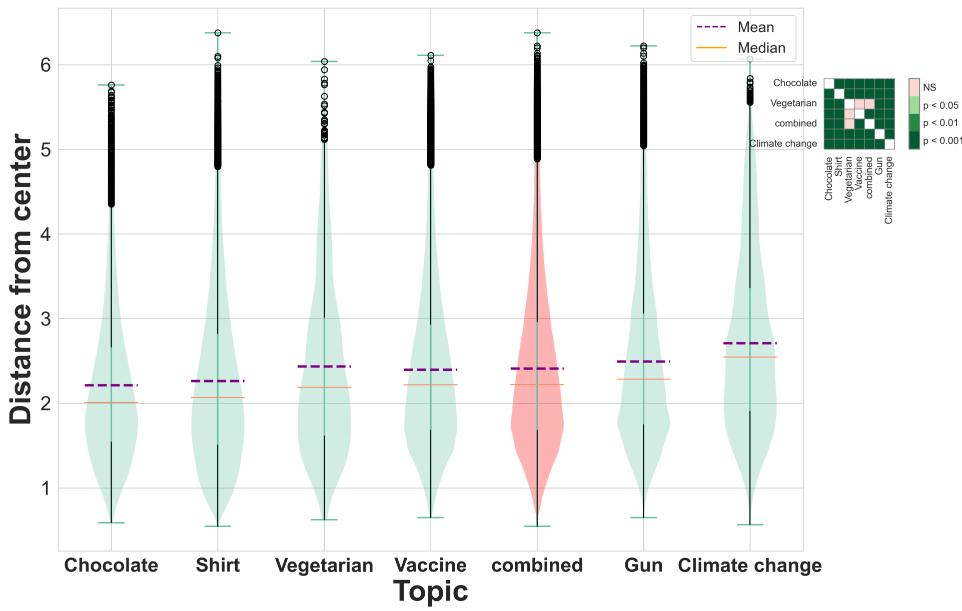

Figure 2: Eccentricity density distribution of opinions: Density distribution of opinion eccentricity is plotted for six topics, with varying controversies. Purple dashed and orange solid line represents mean and median of each distribution, respectively. The red violin plot represents the eccentricity distribution of combined data serving as the baseline for controversial topic identification. Plots are in the order of median eccentricity value of the topic (left plot with the smallest median while the right plot with highest median value)

5 Results

The first key observation we make in our study is that the eccentricity of all opinions in an opinion pool collectively characterizes the diversity of the pool. To better understand, if all the opinions in the pool are mostly homogenous, eccentricity, i.e., the distance from the center, of these opinions is significantly small. However, if we have diverse opinions in the pool, most of them will lie far from the center with higher eccentricity value. Hence, if the density of 'highly eccentric opinions' for a topic is high, it indicates that the topic is more diversified and vice versa.

Interestingly, this leads to our primary finding that the collective eccentricity or diversity of a topic is related with the controversiality of the topic, making it a measure of controversiality. Figure 2 shows the distribution of posts' eccentricity as violin plots; each plot represents distribution of eccentricity for a different topic. The least controversial topics such as Chocolate and Shirt, have the smallest mean and median value for eccentricities. However, the violins for the topics Gun and Climate change, which are strongly controversial, are shifted upwards, with significantly higher mean and median indicating higher density of more eccentric opinions. This observation shows an association between diversity and controversiality of a topic. The red violin plot represents the eccentricity distribution of combined data.

This observation is more evident in Figure 3, when we plot a cumulative distribution of eccentricity for each topic. The curves for the least controversial topics are shifted towards the left while highly controversial topic curves are located on the right. We exploit this relation to quantify the controversiality of a topic in the following ways. We propose to use the median value of eccentricity distribution of a topic as the controversiality metric of that topic. For example, from Figure 2, the controversiality of chocolate is 2.00 while the controversiality of climate change is 2.55. Using median value as the controversiality measure is advantageous since it is not affected by the outliers and makes it easy to compare the controversiality of different topics. We also propose a more relaxed way of quantifying controversiality range for a topic instead of a single value by utilizing the distribution quantiles. We use the range from eccentricity value at 40th quantile to eccentricity value at 60th quantile as the controversiality range of the topic. For example, from Figure 2, the controversiality of climate change is 2.28 - 2.82 while for chocolate it is 1.81 – 2.23.

The proposed measure gives a better understanding of controversiality without providing a false interpretation while comparing controversialities of different topics. For example, from the Figure 2, if we compare the controversiality of Vaccine and Vegetarian using the median method, the Vaccine is more eccentric than the Vegetarian (2.21 vs. 2.19), however, if we use any other measures of central tendency like mean or any quantile range the relation gets inverted. That suggests that in case of quantifying controversiality of topic which are very similar in their controversiality, median can provide a false understanding. But if the topics are well separated from each other in terms of controversiality, using median is appropriate and easier. We can also use a wider/narrower quantile range or any other measure of central tendency of distribution depending upon the requirement and dataset.

Another key observation we make in our study is that the trend of eccentricity density distribution for several topics, that differ in terms of their controversiality, are similar. In Figure 3, though the controversialities are significantly different for topics like Chocolate - Gun or Shirt - Climate, the pattern of density distribution is quite similar. To further examine this observation, we try to fit a common mathematical function to these cumulative density curves. Figure 4 shows the results of curve fitting where each distribution is fitted to a sigmoid function ($Y = \frac{1}{1+e^{-k(x-x_0)}} + b$) with appropriate parameter values. The quality of fit for each

DOI: 10.22191/nejcs/vol5/iss1/2

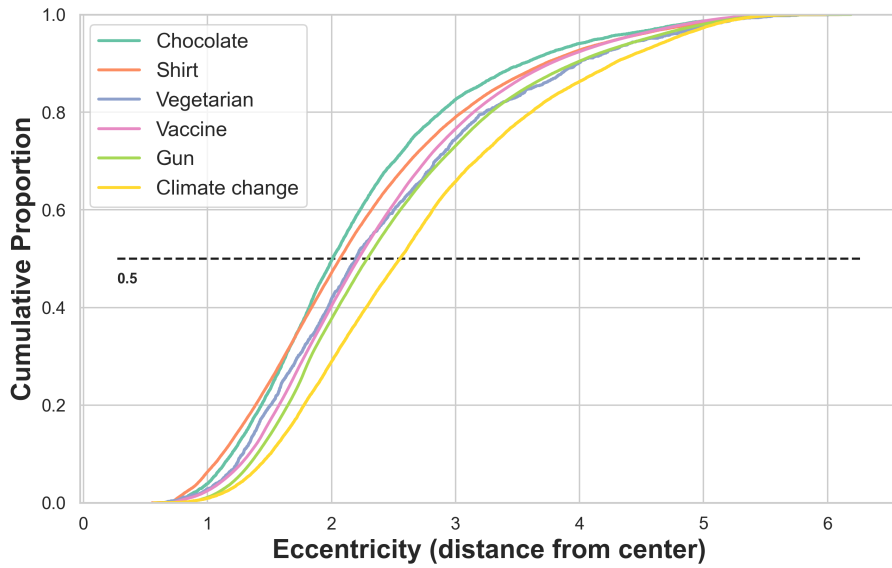

Figure 3: Cumulative distribution of eccentricities for topic with varying controversialities: The topics having different controversiality show similar diversity pattern.

curve is shown in the figure. This suggests that for several topics people frequently express opinions which are far from the norm of society irrespective of the controversiality. This finding explains the trend we are observing in society nowadays when people are opting for extreme choices in their routine life and topics like food, fashion, entertainment which are not traditionally controversial otherwise.

6 Conclusion

In this study, we focused on analyzing data from Parler, a social network known for its highly homogenous and opinionated user base. We demonstrated that the collective eccentricity of opinions on a topic can be used as a measure of the topic's diversity and controversiality. Specifically, we proposed two methods for utilizing the relationship between controversiality and collective eccentricity distribution, highlighting scenarios where each method may be preferred. Our analysis also revealed a striking similarity in the distribution of opinion eccentricity across topics with varying levels of controversiality. This observation has implications beyond our study and suggests that the distribution of opinion eccentricity is a common pattern in routine life. Furthermore, our results imply that if highly homogenous communities such as Parler exhibit a wide range of opinions on a topic, we can expect a more

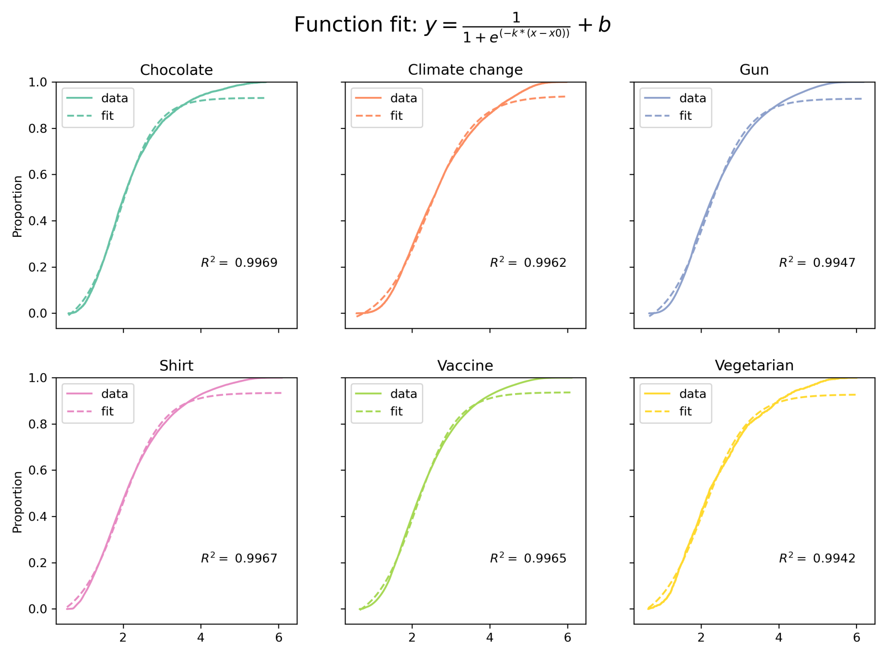

Figure 4: Eccentricity density distribution trend: Independent of the controversiality, all topics follow similar eccentricity density trend. The cumulative density distribution of all topics is fitted to a sigmoid shaped function $(Y = 1/(1 + e^{(-k(x-x_0))}) + b$ with appropriate parameter values. The solid line in each subplot represents the actual cumulative density plot while the dashed line shows the fitted function. The value of goodness of fit R^2 for all the curves is greater than 0.99.

diverse range of opinions in mainstream social media platforms like Twitter or Reddit. Moving forward, we plan to extend our research by analyzing data from other communities with varying levels of homogeneity and heterogeneity to strengthen our findings. Additionally, we aim to explore how social neighborhoods may influence the relationship between controversiality and collective eccentricity. Our hope is that these future directions will deepen our understanding of how opinions are formed and shaped in social networks.

References

[1] Adamic, L. A. and Glance, N. "The political blogosphere and the 2004 U.S. election: divided they blog". In: *Proceedings of the 3rd international workshop on Link discovery*. LinkKDD '05. New York, NY, USA: Association for Computing Machinery, Aug. 2005, pp. 36–43. ISBN: 978-1-59593-215-0.

[2] Akoglu, L. "Quantifying Political Polarity Based on Bipartite Opinion Networks". en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1 (May 2014). Number: 1, pp. 2–11. ISSN: 2334-0770.

[3] Al Amin, M. T. et al. "Unveiling polarization in social networks: A matrix factorization approach". In: *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*. May 2017, pp. 1–9.

[4] Aliapoulios, M. et al. "A Large Open Dataset from the Parler Social Network". en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 15 (May 2021), pp. 943–951. ISSN: 2334-0770.

[5] Aliapoulios, M. et al. *An Early Look at the Parler Online Social Network*. arXiv:2101.03820 [physics]. Feb. 2021.

[6] Al-Ayyoub, M. et al. "Studying the controversy in online crowds' interactions". In: *Applied Soft Computing* 66.C (May 2018), pp. 557–563. ISSN: 1568-4946.

[7] Braman, D. and Kahan, D. M. "Overcoming the Fear of Guns, the Fear of Gun Control, and the Fear of Cultural Politics: Constructing a Better Gun Debate". In: *Emory Law Journal* 55 (2006), p. 569.

[8] Can, U. and Alatas, B. "A new direction in social network analysis: Online social network analysis problems and applications". en. In: *Physica A: Statistical Mechanics and its Applications* 535 (Dec. 2019), p. 122372. ISSN: 0378-4371.

[9] Choi, Y., Jung, Y., and Myaeng, S.-H. "Identifying Controversial Issues and Their Sub-topics in News Articles". en. In: *Intelligence and Security Informatics*. Ed. by H. Chen et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2010, pp. 140–153. ISBN: 978-3-642-13601-6.

[10] Cinelli, M. et al. "The echo chamber effect on social media". In: *Proceedings of the National Academy of Sciences* 118.9 (Mar. 2021). Publisher: Proceedings of the National Academy of Sciences, e2023301118.

[11] Conover, M. et al. "Political Polarization on Twitter". en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 5.1 (2011). Number: 1, pp. 89–96. ISSN: 2334-0770.

[12] Cook, E. A. *: Public Opinion And The Politics Of Abortion*. New York: Routledge, Dec. 2019. ISBN: 978-0-429-04231-7.

[13] Dori-Hacohen, S. and Allan, J. "Automated Controversy Detection on the Web". en. In: *Advances in Information Retrieval*. Ed. by A. Hanbury et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 423–434. ISBN: 978-3-319-16354-3.

[14] Garimella, K. et al. "Quantifying Controversy on Social Media". In: *ACM Transactions on Social Computing* 1.1 (Jan. 2018), 3:1–3:27. ISSN: 2469-7818.

[15] Garimella, K. et al. "Reducing Controversy by Connecting Opposing Views". In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. WSDM '17. New York, NY, USA: Association for Computing Machinery, Feb. 2017, pp. 81–90. ISBN: 978-1-4503-4675-7.

[16] Guerra, P. et al. "A Measure of Polarization on Social Media Networks Based on Community Boundaries". en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 7.1 (2013). Number: 1, pp. 215–224. ISSN: 2334-0770.

[17] Jang, M. et al. "Probabilistic Approaches to Controversy Detection". In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. CIKM '16. New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 2069–2072. ISBN: 978-1-4503-4073-1.

[18] Le, Q. and Mikolov, T. "Distributed Representations of Sentences and Documents". en. In: *Proceedings of the 31st International Conference on Machine Learning*. ISSN: 1938-7228. PMLR, June 2014, pp. 1188–1196.

[19] Mikolov, T. et al. *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781 [cs]. Sept. 2013.

[20] Morales, A. J. et al. “Measuring political polarization: Twitter shows the two sides of Venezuela”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25.3 (Mar. 2015). Publisher: American Institute of Physics, p. 033114. ISSN: 1054-1500.

[21] Popescu, A.-M. and Pennacchiotti, M. “Detecting controversial events from twitter”. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. CIKM ’10. New York, NY, USA: Association for Computing Machinery, Oct. 2010, pp. 1873–1876. ISBN: 978-1-4503-0099-5.

[22] Qiu, J., Lin, Z., and Shuai, Q. “Investigating the opinions distribution in the controversy on social media”. en. In: *Information Sciences* 489 (July 2019), pp. 274–288. ISSN: 0020-0255.

[23] Sullins, D. P. “Catholic/Protestant Trends on Abortion: Convergence and Polarity”. In: *Journal for the Scientific Study of Religion* 38.3 (1999). Publisher: [Society for the Scientific Study of Religion, Wiley], pp. 354–369. ISSN: 0021-8294.

[24] Temperman, J. “Freedom of Expression and Religious Sensitivities in Pluralist Societies: Facing the Challenge of Extreme Speech”. In: *Brigham Young University Law Review* 2011 (2011), p. 729.

[25] Tsytsarau, M., Palpanas, T., and Denecke, K. “Scalable Detection of Sentiment-Based Contradictions”. In: *CEUR Workshop Proceedings* 762 (Jan. 2011).

[26] Zarate, J. M. O. de et al. “Measuring Controversy in Social Networks Through NLP”. In: *String Processing and Information Retrieval: 27th International Symposium, SPIRE 2020, Orlando, FL, USA, October 13–15, 2020, Proceedings*. Berlin, Heidelberg: Springer-Verlag, Oct. 2020, pp. 194–209. ISBN: 978-3-030-59211-0.