Community Guidelines: How Misinformation Reinforces Insular, Polarized, and Toxic Interactions

HANS W. A. HANLEY, Stanford University, USA
ZAKIR DURUMERIC, Stanford University, USA

How do different communities respond to news from unreliable sources? How does news from these types of sources change and alter conversations within these communities? In this work, we examine the role of online misinformation in sparking political incivility and toxicity on the social media platform Reddit's online communities or subreddits. Utilizing the Google Jigsaw Perspective API to identify online examples of online toxicity, hate speech, and other forms of incivility, we find Reddit comments on misinformation articles promote toxicity at a rate 60% higher than on authentic news articles. Identifying the specific instances of commenter's incivility and utilizing an exponential random graph model, we further identify that misinformation, in particular, promotes inter-political strife among Reddit users, with users in these environments more likely to target with toxic comments other Reddit users of different political beliefs than in other settings. Finally, utilizing a zero-inflated negative binomial regression, we identify the specific characteristics of subreddits that promote user engagement with misinformation. In contrast to user engagement with authentic news, users in more uncivil/toxic subreddits are more likely to comment and otherwise engage with misinformation.

CCS Concepts: • **Human-centered computing** → *Collaborative and social computing*; **Empirical studies in collaborative and social computing**; • **Information systems** → **Web Mining**; • **Networks** → *Online social networks*;

Additional Key Words and Phrases: Misinformation, Toxicity, Political Polarization, Reddit, Online Communities

ACM Reference Format:
Hans W. A. Hanley and Zakir Durumeric2022. Community Guidelines: How Misinformation Reinforces Insular, Polarized, and Toxic Interactions. 1, 1 (September 2022), 31 pages.

1 INTRODUCTION

Over the last few decades, misinformation, incivility, and political polarization have corroded trust in democratic institutions [17, 23, 47, 48, 51]. Going viral on social media in the run-up to his 2022 run for a Missouri US Senate seat, former governor, Eric Greitens released a political advertisement calling for primary voters to order a "RINO (Republican in name only) hunting permit." Roundly criticized for promoting violence [37], the popularity of the advertisement on social media illustrated the combination of misinformation, online toxic behaviors, and political polarization that have plagued many social media platforms [27, 48, 91, 107].

While separate and distinct phenomena, misinformation, toxic language, and political polarization are all factors that often combine with one another, stoking division and negatively affecting social media platforms [19, 27, 31, 33, 50, 79, 101, 107, 111]. While several works have attempted to understand the impact of these individual factors on social media, in this work, we explore them in tandem. Specifically, we seek to understand how misinformation promotes uncivil and political charged conversations across different types of online communities, asking the following three research questions:

Authors' addresses: Hans W. A. Hanley, hhanley@stanford.edu, Stanford University, 450 Serra Mall, Stanford, California, USA, 94305; Zakir Durumeric, zakird@stanford.edu, Stanford University, 450 Serra Mall, Stanford, California, USA, 94305.

(1) *How do online toxicity and political polarization norms in a given community correlate with the presence of misinformation and unreliable sources?*
(2) *Does political polarization inform why users across different communities are uncivil or toxic with one another? How is this exacerbated or affected by the presence of misinformation?*
(3) *Do online toxicity and political polarization community norms affect how and when users interact with misinformation and unreliable news?*

To answer these questions, we examine the submissions and comments of 18 months of Reddit data from June 2020 to July 2021. Using the Google Jigsaw Perspective API [2], an out-of-the-box classifier for identifying uncivil and toxic language, we determine the relative rates at which toxic comments are published within each subreddit. Utilizing a domain-based approach as outlined by Saveski et al. [96], we next approximate the political orientations of a subset of Reddit communities along the US left-right political spectrum. Having calculated rough outlines of toxicity and political norms within each subreddit, we finally determine how these levels of toxicity and political orientations influence how users interact with each other when influenced and when not influenced by misinformation.

RQ1: Toxicity and political polarization in Reddit misinformation posts. Reddit enables users to submit news articles as submissions to different communities. Underneath these submissions, users can leave comments on the article or on the comments of different users. Utilizing two lists of misinformation and authentic news websites, we thus first determine the levels of political polarization and toxicity in the comments of Reddit submissions of news articles from misinformation and authentic news sources. Looking at these comments, we find commenters on news articles from misinformation are toxic at a rate 63% higher than commenters on authentic news (1.17% of misinformation comments versus 0.7% of authentic news comments). We similarly find that users post toxic comments at a rate between 57% and 75% higher in subreddits where misinformation news articles are posted compared to those where authentic news articles are posted. Looking at the approximate political polarization levels of the commenters on misinformation and authentic news, we find that they come from across the political aisle for both misinformation and authentic news submissions. However, looking at the difference in the distribution of political views of those that post misinformation versus those that comment, we find that the posters are on the whole more conservative than their corresponding liberal commenters.

RQ2: Misinformation fueling inter-political strife. Having identified that misinformation commenters are more toxic than authentic news commenters, we next determine the role of a key driver of this increased incivility: political polarization. Specifically, we find that users that post underneath misinformation submissions are more likely to be toxic to other users of different political beliefs than those of similar political views (odds ratio 1.63). Utilizing an exponential random graph model to confirm this finding, we show that indeed, for misinformation commenters, the more different they are from other users, the more likely they are to respond to them in a toxic manner. By approximating the levels of misinformation within subreddits in our dataset, we find that as subreddits have more misinformation-related content posted within them, the more likely users of different political orientations are to be toxic with each other. We specifically see the rate at which misinformation induces inter-political toxicity is nearly 2.3 times the rate at which it drives intra-political toxicity.

RQ3: Toxic subreddits powering engagement with misinformation Lastly having documented the role of misinformation in promoting toxicity, especially among users of different political orientations, we determine how different levels of user toxicity and polarization levels affect community engagement with misinformation compared with authentic news. Fitting a zero-inflated negative binomial model to our data, we find, specifically, that as subreddits become more

toxic and politically polarized, their users are more likely to engage with misinformation. This is in contrast to authentic news, where more toxic communities are less likely to engage with their articles.

Altogether, in this work, we document the role that misinformation has in creating more politically insular and toxic communities. We find within communities across Reddit that levels of misinformation are correlated with toxicity and inter-political strife. Our work, one of the first to examine the relationship between misinformation, toxicity, and political polarization illustrates the need to fully understand the effect of misinformation across different platforms. We hope that this work helps inform other research into how misinformation and unreliable sources negatively affect the health of online communities.

2 BACKGROUND & RELATED WORK

Increasingly the role of social media in promoting misinformation-heavy, toxic, and highly politically polarized ecosystems has been intensely studied [26, 53, 57, 105]. In this section, we detail several key definitions that we utilize to operationalize our study, give background on Reddit, and finally present an overview of existing works that scrutinize the effects of misinformation, toxicity, and polarization factors on different social media platforms.

2.1 Terminology

We first provide several key definitions that are core to our work.

Misinformation, Unreliable Sources, and Authentic News: As in previous works, we define *misinformation* as information that is false or inaccurate regardless of the intention of the author [8, 54, 57, 64, 69, 76, 115]. Similarly, as in several previous works [4, 8, 24, 57, 61, 84, 99, 122], we define *misinformation websites* or "unreliable sources" as news websites that regularly publish false information or misinformation about current events and that do not engage in journalistic norms such as attributing authors and correcting errors. As in these past works, we specifically choose to utilize a definition of *misinformation websites* that encompasses websites of different types that regularly publish false information. Conversely, we further define *authentic news websites* as news websites that generally adhere to journalistic norms including attributing authors and correcting errors; altogether publishing mostly true information [57, 61, 122].

To operationalize and understand how misinformation affects different Reddit users and communities, we examine how they interact with news articles and information from *misinformation websites* compared to other *authentic* and reliable sources. In this way, we seek to understand how Reddit users interact with news from largely unreliable sources and in environments where these types of sources are heavily utilized.

Online Toxicity and Incivility: Given our use of the Perspective API [2], we define online toxicity and incivility as it does; namely toxicity is: "*(explicit) rudeness, disrespect or unreasonableness of a comment that is likely to make one leave the discussion.*" Within the conversations that we analyze that center around news and misinformation, we thus consider comments that meet this definition to be toxic/uncivil.

Political Polarization and Orientation: We define political polarization/orientation as users and communities place on the US left/right political spectrum [93]. We note the limitation of this definition given the variety of political views within the US. However, in line with previous work [56, 95, 96], we utilize this definition to understand how conservative-leaning and liberal-leaning users and communities interact with one another and misinformation.

2.2 Reddit

As previously stated, Reddit is an online website composed of millions of different subcommunities known as subreddits [3, 21]. These subreddits are each dedicated to specific topics, ranging from politics (r/politics) to science (r/science) to memes (r/memes). Reddit enables users to submit news articles, opinions, images, and memes as submissions within subreddits. Underneath these submissions, users can leave comments on the submission or on the comments of different users. Subreddits can be created by anyone and they are moderated both by Reddit content policies as well as subreddit-specific rules and often implicit community guidelines [21, 38, 68]. As discussed by Chandrasekharan *et al.*, subreddits also have various macro-, mes-, and micro-norms that govern behavior [21]. These norms extend to political behaviors, tolerance of misinformation, and toxicity [21, 68, 91, 116]. Weld *et al.* [116] find that these norms vary widely, with each subreddit having its own unique value hierarchy. In this work, we investigate how these norms inform online conversations.

2.3 Political Polarization

People, both in real life and on the Internet, tend to associate and be friends with like-minded people [12, 13, 55, 65, 70, 88]. While social media can have the benefit of exposing individuals to multiple views and allow them to interact with different types of people [13, 32, 88], many studies find it to be one of the sources of the high degree of polarization in various countries [19, 20, 59, 70]. Cass Sunstein, Garett *et al.*, and Quattrociocchi *et al.* all argue that the "individualized" experience offered by social media companies comes with the risk of creating "information cocoons" and "echo chambers" that accelerate polarization [46, 89, 104]. Several other authors have further interrogated the negative effects that social media has had on the democratic process due to the levels of polarization systematically promoted by social media [52, 86, 107, 108]. In addition to polarization being amplified by social media, other works have found this increased polarization to cause upticks in the amount of misinformation and toxic behaviors online. Imhoff *et al.* [67], for example, find that political polarization, on both sides of the political spectrum, is associated with beliefs in conspiracy theories; Hanley *et al.* [56] confirm this with their own Internet-wide study. Ebling *et al.* [35], find that political partisanship levels on social media are associated with medical misinformation about COVID-19.

As a result of these concerns, various works have sought to understand the role of political polarization on social media. Several works have found that social media does indeed create political echo-chambers [102], where users' biases are reconfirmed and reinforced [9, 16, 26, 29]. Bessi *et al.* [16], examining the behaviors of over 12 million users, find that partisan echo chambers were driven by the algorithms of both Facebook and YouTube. These echo chambers can often lead to heavily partisan users. For instance, Torres *et al.* [106] find the specific Twitter behavior of "follow trains" induced highly politically polarized behavior on the platform. Finally Conover *et al.* [29], find that different structures of Twitter interactions as a function of how the platform was built, foster different levels of politically polarized conversations. Wojcieszak *et al.* [118] find that the majority of political discussions online are between participants that share the same viewpoint, this is even though many users *do* enjoy conversations with people with different viewpoints [103]. An *et al.* for example, find that users who got their information from partisan social media had "distorted views of reality" [9].

2.4 Misinformation

In addition to powering heavily polarized users, online activity has been found to be the main driver for the spread of misinformation. As researched and reported extensively, misinformation has

increasingly become a major and distinctive aspect of the information on social media [8, 44, 49]. Vousoughi *et al.* [112] even find that on Twitter, misinformation spreads 10 times faster than true information. Furthermore, misinformation often convinces those that are exposed to it. A large percentage of US adults were exposed to misinformation stories by social media during the 2016 election [8] and many believed these false stories were true [7, 54]. As COVID-19 spread throughout the world, misinformation and conspiracy theories became a major hurdle to curbing its spread [94, 100].

To prevent the spread of misinformation, research has heavily focused on tracking and stemming its flow[57, 107]. Mahl *et al.* [78], track the spread of 10 different conspiracy theories on Twitter, identifying one of the largest conspiracy theorist networks. Ahmed *et al.* [5] use a similar approach to track the spread of COVID-19 and 5G conspiracy theories. They find well-known misinformation websites were some of the largest sources helping to spread these conspiracy theories on Twitter's platform. In this same vein, Gruzd [53] found that a single Tweet about how COVID-19 was a hoax, spanned an entire conspiracy theory sending large groups of people to film their local hospitals to prove that COVID-19 was not real. Wood *et al.* [119] utilized similar social network analyses to understand conspiracy theories that were spread on Twitter about the Zika virus outbreak. In addition, to utilize network analysis, several other works take a natural language processing-based approach. Hanley *et al.* [58] for example, utilize semantic search to identity and track Russian state-media narratives on Reddit. Fong *et al.* [40] utilized linguistic and social features to understand the psychology of Twitter users that engaged with known conspiracy theorists on the Twitter platform. Finally, several works have performed in-depth case studies on the spread of specific misinformation narratives. In their papers, Wilson and Starbird *et al.* look at the Syrian White Helmets on Twitter, Papadamou *et al.* look at disturbing videos targeting children on YouTube, and Bär *et al.* look at the spread of QAnon on Parler [18, 83, 117].

2.5 Toxicity

41% of Americans and 40% of those globally have reported experiencing bullying or harassment online [34, 105]. Online toxicity takes many forms including intimate partner violence, sexual harassment, doxing, cyberstalking, coordinated bullying, political incivility, and account takeovers [42, 43, 77, 105]. In addition to our definition, Vargo *et al.* [110] describe toxic comments as those that utilize "extremely vulgar, abusive, or hurtful language". Similarly Muddiman *et al.* define online political toxicity [81] as comments that violate "politeness norms, such as name-calling and swearing, and democratic norms, such as claims of discrimination, government dysfunction, and treason."

Toxic comments are one of the most common forms of hate and harassment online [105]. Toxicity is also a key aspect of social media [30, 73, 82, 105, 120]. Facebook estimates that between 0.14% and 0.15% of all views on their platform are of toxic comments [36]. Saveski *et al.* [97] further find that particular structures of Twitter conversations lead to tweets with toxic language. This type of incivility, in addition to damaging online conversations, has been found to also damage civil institutions [17, 108] having dangerous real-world implications. Fink *et al.* [39] find that politically charged anti-Muslim hate speech on Facebook in Myanmar was a prominent aspect preceding the Rohingya genocide.

To prevent the spread of toxic content, various platforms have implemented and designed a variety of safeguards [1, 2, 36]. Researchers have further taken to performing in-depth studies on users' behavior to understand abusers and victims of abuse. For instance, Founta *et al.* [41] identify a set of network and account characteristics of abusive accounts on Twitter. Hua *et al.* [63] look at properties of the accounts that have heavily negative interactions with political candidates on

Twitter. Finally Chang *et al.*, Xia *et al.*, Zhang *et al.*, and Lambert *et al.* all look at the set of causes that make conversations unhealthy or toxic [75, 121, 123, 124].

2.6 The Interplay of Misinformation, Online Toxicity, and Political Polarization

Several works, close to our study, have attempted to understand how political polarization, online toxicity, and misinformation interact. Online toxicity, for instance, has been heavily associated with increased political polarization and the use of misinformation [27, 107]. Cinelli *et al.* [27], in particular, find that misinformation about COVID-19 on YouTube ended up promoting hate, toxicity, and conspiracy theories on the platform. Similarly Chen *et al.* [23] utilize network analysis to find that misleading online videos lead to increased incivility online.

In a separate vein, Rains *et al.* [90] find that high polarization is a major factor in producing incivility and toxicity online. Bail *et al.* [10] and De Francisci Morales *et al.* [32] find, most markedly that the interaction of individuals of different political orientations increased toxicity and negative conversational outcomes. Similarly, Kim *et al.*, Kwon *et al.*, and Shen *et al.* all find that exposure to these negative conversations actually increases observers' tendency to also engage in incivility [72, 74, 101]. Finally, linking misinformation to political polarization, Hanley *et al.* [56] and Imhoff *et al.* [67] find that political polarization is a key aspect of people's belief in false narratives.

3 DATASETS & METHODS

Within this work, we study levels and the interaction of misinformation, toxicity, political polarization on Reddit. Many previous works have individually investigated each of these topics thoroughly on these platforms as well as the wider Internet; we thus rely on previous work when compiling our datasets. In addition to giving an overview of these datasets in this section, we give a brief overview of how we calculate the partisan bias of social media users, how we gather the toxicity levels of tweets and comments, and finally how we determine levels of misinformation within different subreddits.

3.1 Reddit Dataset

In order to model the impact of misinformation relative to the political and toxicity community norms within different social media communities, we utilized 18 months of Reddit comments and submissions from January 2020 to June 2021. In order to collect this data, we rely on the Pushshift [15], a third-party API that collects and publishes monthly datasets of Reddit data. Each comment and submission includes a timestamp, the author's username, the subreddit/community where the comment was posted, as well as the particular conversation thread where the comment was posted. Using this data, we reconstruct the different threads and posting behaviors for each user and each subreddit. Given our focus on English-language misinformation websites and the Perspective API to estimate toxicity, we further filter our dataset to include only English language comments and submissions using the `whatlanggo` Go library.[1]

Approximating the Political Polarization of Subreddits and Users. To approximate the partisanship or political polarization of different subreddits and users, we utilize a dataset of website partisanship scores developed by Robertson *et al.* [93]. Robertson *et al.*'s original dataset measured the partisanship of different sites based on how often they were shared by Democrats and Republicans on Twitter in late 2017. Their dataset includes partisan bias scores for 19K websites, giving each a score between -1 (liberal/Democratic-leaning) and +1 (conservative/Republican-leaning).

To estimate the approximate political leaning of subreddits and users, we take the average of the political partisanship scores of the hyperlinks that they posted online. For example, if a user posts

[1] https://github.com/abadojack/whatlanggo

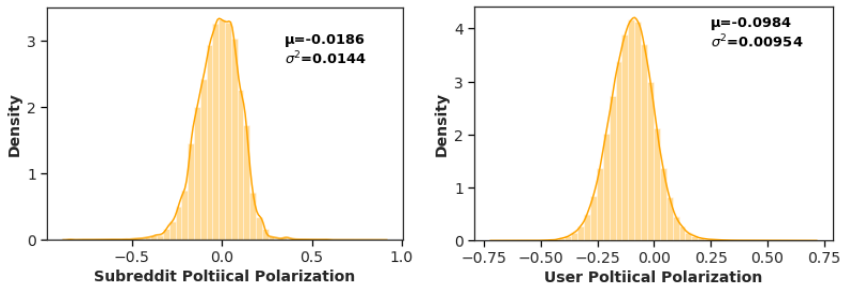


Fig. 1. **Distribution of subreddit and user polarization scores** — We estimate the political polarization of users and subreddits based on the political polarization of the URLs they post. We compute these estimates for users and subreddits that have posted at least 10 URLs to get robust averages for each subreddit and user. Altogether we get approximate political polarization scores for 427K users and 46,681 subreddits.

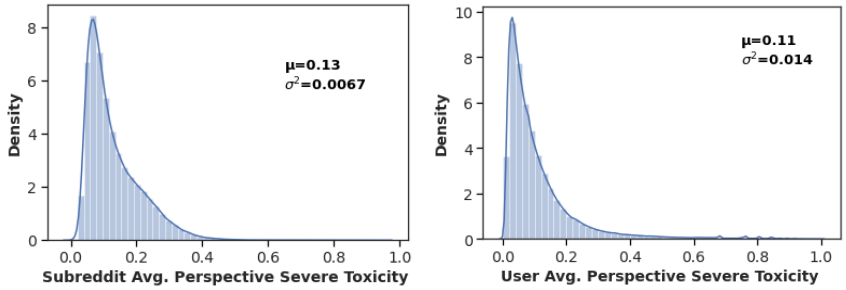
Conservative Subreddits	Polarization Score	Liberal Subreddits	Polarization Score
r/The_Generals_Corner	0.876	r/Illinois4Sanders	-0.843
r/AmericanThinkerForum	0.829	r/Georgia4Sanders	-0.843
r/PhillyGuns	0.782	r/NewMexicoForSanders	-0.843
r/FightingFakeNews	0.758	r/alaskaforsanders	-0.843
r/HatedForNoReason777	0.742	r/dc4sanders	-0.843

Table 1. **The most politically charged subreddits**— The most liberal subreddits are those advocating for US Senator Bernie Sanders.

frequently from both nytimes.com (-0.2602 Democratic/Liberal) and veteranstoday.com(+0.2994 Republican/Conservative), this would result in an approximate political partisanship score of 0.0392. As found by Saveski *et al.* [95], utilizing the polarization of URLs posted by users was found to largely correlate (R^2 = 0.81) with users' US voting behaviors. We thus utilize these scores to estimate subreddits' and users' political leaning on the US political spectrum. We further note that while these subreddits and users may not be overtly political, their use of politically charged and biased URLs does allow us as in Saveski *et al.* [95] to approximate their political leanings.

To get a robust score for each user and subreddit, we only utilize averaged scores for users and subreddits who have posted more than 10 URLs. Furthermore, we note, that to approximate user political leanings we utilize all URLs posted by the user both in their Reddit submissions as well as their comments. In contrast, we only utilize the URLs posted in submissions on subreddits when calculating their political leanings. We do this as these hyperlinks are implicitly approved by the given subreddit community and are more reflective of the political leanings of the full subreddit [114]. We further remove internal Reddit hyperlinks when calculating the political leaning of users (*i.e* a Reddit user or subreddit hyperlinking to another page on Reddit does not affect the political leaning calculation.) Altogether, we calculate and utilize scores for 427K users and 46,681 subreddits.

As seen in Figure 1, the average political leaning of Reddit users is liberal/Democratic-leaning. This largely agrees with Pew Research polling data which found that 47% of Reddit users identify as liberal, 39% as moderate, and 13% as conservative [14]. In contrast, we see across our measured subreddits, that the average is only slightly liberal-leaning. This indicates that while subreddits are



caption**Distribution of subreddit average and user average Perspective Severe Toxicity scores** — We determine the toxicity norms for subreddits with at least 50 comments and users with at least 10 comments. Each user and subreddit has distinctive toxicity norms, posting toxic comments at different rates. At a threshold of 0.8, most users and subreddit's usual comments/posts are not considered toxic or pernicious by the Perspective API SEVERE_TOXICITY classifier.

Most Toxic Subreddits	Comment Toxic %
r/141414	100%
r/ElectionPollsUSA	100%
r/spamswearshere	100%
r/icummedtothat	88.4%
r/WorshipSofiaVergara	81.5%

Table 2. **The Most Toxic Subreddits**

created for individuals across the political spectrum [91], the liberal/Democratic-leaning subreddits are the most popular and have the most users.

Identifying Toxic Reddit Comments and Approximating Users and Subreddit Toxicity Norms. To approximate the relative toxicity of Reddit users and subreddits, we utilize the Perspective API, a set of out-of-of-the-box toxicity classifiers from Google Jigsaw [2]. The Perspective API takes comments as input and returns a score of 0–1 for several different classifiers. For each classifier, the closer that each comment's score is to 1, the more likely the comment is pernicious or toxic.

The Perspective API has been utilized extensively in prior works [73, 91, 97] and we rely on the best practices outlined in past works for our study. For this work, to precisely pinpoint explicit examples of highly toxic comments, we specifically utilize the SEVERE_TOXICITY classifier. As in Chong *et al.* and other works, to consider a comment as toxic, we utilize a threshold of 0.8 [25, 75]. As found by Kumar *et al.* [73], utilizing this particular classifier, while limiting recall, provides an acceptable precision at identify toxic online content. Following best practices, we further filter out comments that were less than 15 characters or more than 300 characters in length [73].

To calculate toxicity norms and identify toxic comments, we first identify 31.1 million users within our set of 46,681 subreddits for which we have political data. We then gather all the comments they posted between January 2020 and June 2021 across every subreddit they posted in. We do this across these users' comments in every subreddit to approximate toxicity norms for their behavior overall across Reddit. For each of the 1.7 billion English-language comments from the 31.1M users, we retrieve the Perspective API SEVERE_TOXICITY classifier score. From a subset of these scores, we further determine the approximate toxicity norms for each of the 46,681 subreddits for which we have political data. To approximate each subreddit's and user's toxicity norms/how often they

post toxic content, as in prior work [91], we determine the percentage of toxic comments (based on our definition of toxic comments). We further note, however, that when utilizing toxicity norms within this work, we only utilize those for subreddits with at least 50 comments and users with at least 10 comments. As seen in Figure ??, while there is a wide range of online toxic behaviors, based on our strict definition of toxicity, most users and subreddits are on average benign in their interactions.

3.2 Misinformation and Authentic News Websites

To analyze how users interact with misinformation on Reddit, we utilize a previously curated list of misinformation websites. Specifically, we utilize the same list of misinformation websites utilized in Hanley *et al.*[56]. This list utilizes domains previously curated by the Columbia Journalism Review,[2] OpenSources,[3] Politifact,[4] Snopes,[5] and Melissa Zimdars.[6]. Hanley *et al.*'s list consists of 541 misinformation websites including the likes of The Conservative Treehouse and Info Wars [56]. Many of these websites have been documented as being a part of toxic political echo chambers [102]. As a control against which to measure the behavior of users and subreddits that interact with misinformation URLs on Reddit, we utilize a separate dataset of *authentic news* websites. Similar to our set of misinformation websites, we utilize a previously curated list of authentic news sites gathered by Hanley *et al* [56]. This list again is built from websites documented OpenSources, Politifact, Snopes, and Melissa Zidmars in addition to a list of several local newspapers. Altogether this list consists of 565 different websites from across the political spectrum including websites like cnn.com and dailywire.com.

In addition, in order to verify some of our initial findings, we rerun several of our experiments with an additional separate list of misinformation and authentic news websites. Specifically, as our second set of misinformation websites, we utilize a set of 932 websites labeled as "questionable sources" by the website Media Bias/Fact Check.[7] Media Bias/Fact Check labels websites as a "questionable source" if it exhibits one of the following "extreme bias, consistent promotion of propaganda/conspiracies, poor or no sourcing of credible information, a complete lack of transparency and/or is fake news." This largely matches our definition of misinformation websites outlined in Section 2.1. This list of websites has been utilized throughout prior works [11, 28]. After removing duplicates from our original list of misinformation websites, we were left with 922 websites. As our second set of authentic news websites, we utilize a set of 1885 news websites labeled as center[8], center-left[9], and center-right[10] by Media Bias/Fact Check. After removing duplicates from our original list of 565 websites, we were left with 1743 websites in our second authentic news dataset.

3.3 Misinfo-Oriented and Mainstream-Oriented Websites

Finally, for the rest of this work, in order to approximate how the amount of misinformation and authentic news within given subreddits affect behavior within given subreddits, we define a class of 157,605 domains that are *misinfo-oriented* and 667,848 that are *mainstream-oriented*. We do this given the relatively small size of our set of misinformation and mainstream websites. By extending

[2] https://iffy.news/index
[3] https://github.com/several27/FakeNewsCorpus
[4] https://www.politifact.com/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/
[5] https://github.com/Aloisius/fake-news
[6] https://library.athenstech.edu/fake
[7] https://mediabiasfactcheck.com/fake-news/
[8] https://mediabiasfactcheck.com/center/
[9] https://mediabiasfactcheck.com/leftcenter/
[10] https://mediabiasfactcheck.com/right-center/

our list with *misinfo-oriented* and *mainstream-oriented*, we manage to better approximate how much *misinfo-oriented* and *mainstream-oriented* materials are within specific subreddits.

As in prior work [56], we define *misinfo-oriented* as websites that have more connections from our set of misinformation websites than from authentic news (i.e., the majority of a site's inward links in a domain-based graph are from misinformation websites). Similarly, we define websites as *mainstream-oriented* websites that have more connections from authentic news websites than from misinformation websites. To determine which websites fall into these definitions, we utilize Common Crawl data[11]–widely considered the most complete publicly available source of web crawl data. For each misinformation and authentic news in our first dataset, we collect the set of their domain's HTML pages that were indexed by Common Crawl before August 2021. For each HTML page indexed by Common Crawl, we parse the HTML and collect hyperlinks to other pages (i.e., HTML <a> tags). Using this approach, we then can determine which misinformation and authentic news websites have hyperlink connections with which websites on the Internet. We then determine which websites hyperlinked by our set of misinformation and authentic news websites are *misinfo-oriented* and *mainstream-oriented*. Altogether we gather the available Common Crawl pages and scrape the HTML for 541 misinformation and 565 authentic news websites in our first URL dataset (we do not do all websites in our dataset given issues with the 100s of TBs required Common Crawl data). Websites that have been widely documented as spreading falsehood and conspiracy theories are included within this list as *misinfo-oriented* including infowars.com and 8kun.top [57, 102]. Conversely, our list of *mainstream-oriented* websites includes reputable sources like nytimes.com and wsj.com.

3.4 Ethical considerations

Within this work, we largely focus on identifying large-scale trends in how different subreddits interact with misinformation, levels of toxicity, and levels of political polarization. While we do calculate toxicity and polarization levels for individual users, we do not contact nor attempt to deanonymize them.

4 RQ1: REDDIT MISINFORMATION SUBMISSIONS VS AUTHENTIC NEWS SUBMISSIONS

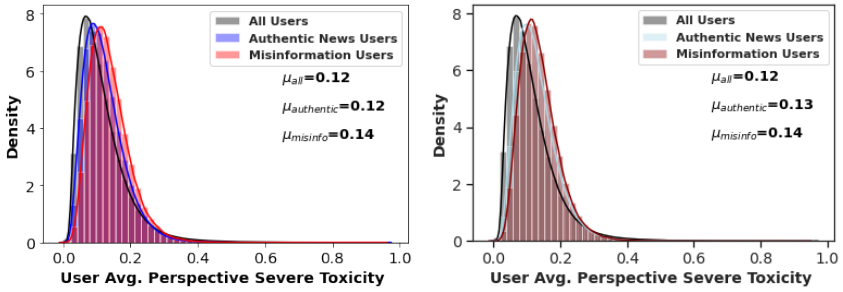
Having outlined our methodology, in this section we turn to understand the relative levels of toxicity and polarization orientations of users and subreddits that interact with misinformation and authentic news.

4.1 Experimental Setup

As previously mentioned, on Reddit, users can submit news articles from different websites as a submission under which users can comment and discuss the article or issues prompted by the news article headline. In order to understand the relative presence of political incivility and toxic comments compelled by misinformation, we thus first compare levels of toxic comments posted under misinformation and authentic news URL submissions.

To do this, across all our collected subreddits we gather the sets the URL submissions that utilize our set of misinformation and authentic news websites. Altogether, within our Pushshift dataset, there were 38,264 different submissions utilizing our first set of 541 misinformation websites from 2,217 different subreddits and 226,930 submissions utilizing our first set of 565 authentic news websites from 18,383 unique subreddits. This difference in magnitude of submissions, we believe, is largely due to the greater popularity and widespread appeal of authentic mainstream news

[11] https://commoncrawl.org/



(a) URL Set 1 (b) URL Set 2

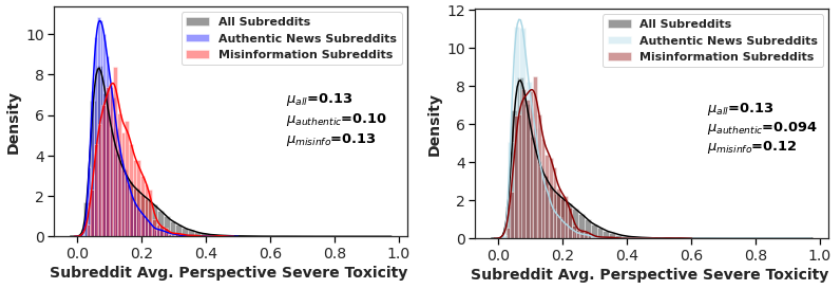
Fig. 2. **Toxicity levels for users who comment under authentic News and misinformation URL Reddit submissions**—Users who interact with misinformation submissions are slightly more toxic/uncivil than users that interact with authentic news. Both groups are slightly more toxic/uncivil than Reddit users generally.

compared with alternative and more fringe websites. Indeed, utilizing the Amazon Alexa Top Million list from March 1, 2021 [6], we find that 255 authentic news websites were in the top 100K websites, while only 101 misinformation websites were in the top 100K. To further bolster and confirms our results, we test our findings in this section utilizing our second set of misinformation and mainstream websites. Altogether, from this second set of URLs, we find an additional set of 9,558 misinformation and 560,673 authentic news submissions.

4.2 Differences in Toxicity/Incivility between Misinformation and Authentic News Submissions

Looking at the toxicity comments from our first set of misinformation domains, we see that 14.9% of the submission had at least one toxic comment and 1.35% of all the comments were toxic. In contrast, for the first set of authentic submissions 13.6% of the submissions had a toxic comment and only 0.85% of the comments were toxic. We thus see a 60% uptick in the rate at which toxic comments are posted on the misinformation submissions. Confirming this finding with our second set of 9,558 misinformation and 560,673 authentic news submissions, we again see a similar pattern of higher toxicity in the misinformation submission comments. 15.3% of the misinformation submissions had toxic comments with 0.92% of the comments being toxic. In contrast, 11.74% of the mainstream submissions had toxic comments with 0.64% of the comments being toxic. We thus see in this replicated experiment that Reddit misinformation conversations indeed have a higher incidence and occurrence of toxicity and incivility. Putting all these conversations together, we find that 1.17% of comments under misinformation submissions are toxic/uncivil while 0.7% of comments for authentic news websites are toxic/uncivil. we thus find an overall 63% increase in the rate at which toxic comments are posted on misinformation submissions.

Higher toxicity within misinformation submissions could be caused by (1) more toxic/uncivil users participating in these conversations or (2) higher toxicity norms in the subreddits where the misinformation was posted. As seen in Figure 2, we see that these users are slightly more toxic than their authentic news counterparts. For the first group of URL submissions, on average 1.54% of the comments for the users associated with these submissions are toxic/uncivil compared with 1.22% for the corresponding group of authentic news users. Similarly, on average 1.48% of comments posted by the second group of misinformation users are toxic compared to 1.32% for the authentic news commenters. However, for bot URL groups, as seen in Figure 2, the distributions of their Perspective API scores are fairly close. We note, that despite this proximity in toxicity of



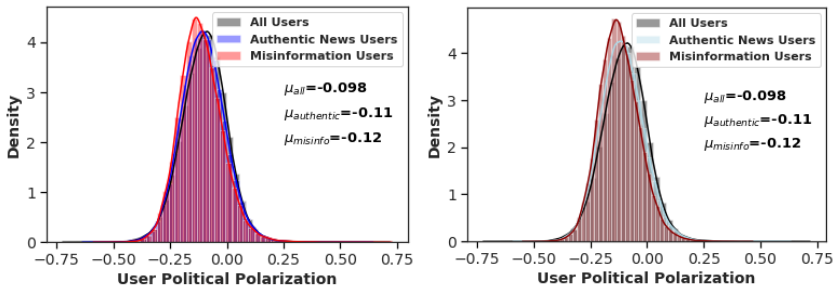
(a) URL Set 1 (b) URL Set 2

Fig. 3. **Toxicity Levels in Subreddits with Authentic News and Misinformation URL Submission**— Subreddits with misinformation submissions are overall more toxic/uncivil compared with authentic news subreddits and subreddits more generally.

misinformation commenters between authentic news commenters, the higher user toxicity appears stable even among users from the same subreddits. Comparing the users who posted in subreddits where *both* mainstream and misinformation URLs were posted, we still see that the users who posted on misinformation submissions had elevated rates of toxicity (1.21% compared to 1.45%). We thus see "more toxic" users are indeed commenting more on misinformation submissions compared to authentic news submissions. Finally, to further confirm our results, we perform U-Mann Whitney tests to ensure that there are indeed statistically significant differences between the rate of toxicity in misinformation and authentic news users; running these tests finding p-values $< 10^{-12}$, we indeed conclude that both groups URL submission commenters that there are indeed higher rates of toxicity for the misinformation users.

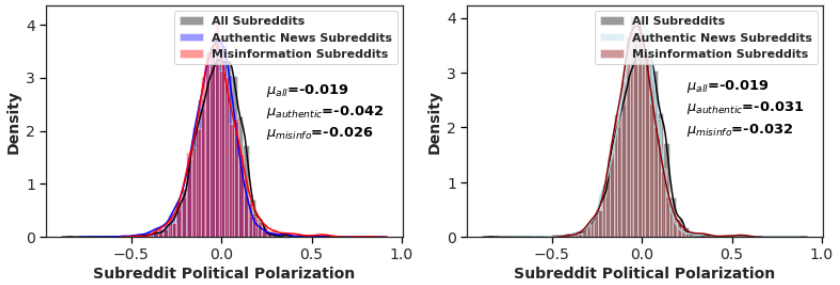
However, despite the finding that more toxic users are indeed commenting more often on misinformation submissions, their higher rate of toxicity is not enough to explain the larger amount of toxic comments in misinformation submissions. After accounting for the higher rate of user toxicity across all the URL submissions, we still see 25.4% more toxic comments than would be expected for the first set of URLs and 28.0% for the second comparison set. Other factors, besides the specific users that comment on misinformation, are contributing to the higher rate of toxicity on misinformation submissions.

Looking at the role of subreddits in promoting toxicity in Figure 5, we find that the toxicity norms of subreddits with misinformation submissions also contribute to higher levels of toxic comments. We see that on average for both sets of URLs that we consider, the set of subreddits with misinformation submissions have higher levels of toxicity compared to subreddits with authentic news submissions. Altogether, in our first set of misinformation URLs, for corresponding subreddits 1.40% of comments posted there are toxic/uncivil. This is compared to a corresponding rate of 0.80% for comments within subreddits with authentic news submissions. We see a similar difference from our second set of URLs (1.1% vs 0.7%). From these two datasets, we thus find that users in subreddits with misinformation submissions post toxic comments at a rate between 57% (second set of URLs) and 75% (first set of URLs) higher than in subreddits with authentic news submissions. We again perform U-Mann Whitney tests to ensure that there are indeed statistically significant differences between the rate of toxicity in misinformation and authentic news subreddits. With p-values $< 10^{-12}$, we conclude that for both groups that there are indeed higher levels of toxicity in the misinformation subreddits.



(a) URL Set 1 (b) URL Set 2

Fig. 4. **Political polarization of users who comment under authentic news and misinformation Reddit submissions**– There are not significant differences in political ideology between users who comment on misinformation and those that comment on authentic news.



(a) URL Set 1 (b) URL Set 2

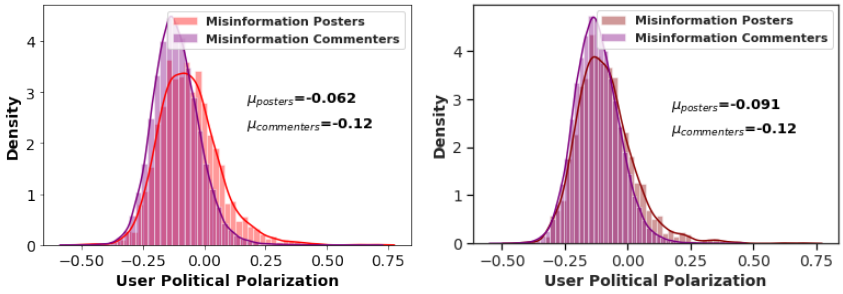
Fig. 5. **Political polarization of subreddits with authentic news and misinformation Reddit submissions**— There are no significant differences in the political orientation of subreddits where misinformation and authentic news appear.

4.3 Differences in Political Polarization between Misinformation and Authentic News Submissions

Having seen the higher levels of toxicity/incivility present within misinformation Reddit submissions, we now explore the differences in political polarization between users who comment on misinformation and those that comment on authentic news.

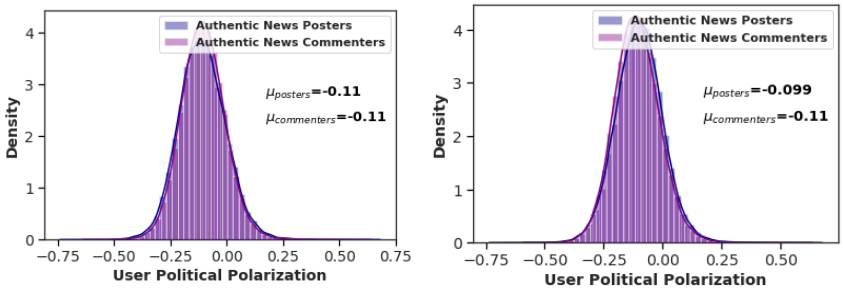
Looking at the set of users commenting under misinformation Reddit submissions, we surprisingly do not see dramatic differences between these users and those that comment on authentic news submissions. In fact, as seen in Figure 4 for both our set of misinformation URL sets, we see a slight leftward tilt in the average commenter. Similarly and surprisingly, looking in Figure 5 at the political orientation of the subreddits where our misinformation submissions appeared, we again see that there is not much difference in their respective polarizations. This appears to indicate that *both* misinformation and authentic news appear within subreddits and get commented on by users across the political spectrum.

We note that despite misinformation appearing in subreddits across the political spectrum, the users that post misinformation have a rightward tilt compared to the users that comment on misinformation. As seen in Figure 6, for both sets of URLs submissions with comments, we see



(a) URL Set 1 (b) URL Set 2

Fig. 6. **Distribution of political orientation of posters and commenters of misinformation**— There is a noticeable rightward tilt in users that post misinformation compared to those that comment on misinformation.



(a) URL Set 1 (b) URL Set 2

Fig. 7. **Distribution of political orientation of posters and commenters of authentic news**— Unlike for misinformation posts, the posters and the commenters on authentic news share similar distributions of political orientations.

that misinformation submitters are on the whole more conservative than their corresponding more liberal commenters. This is largely in contrast to authentic news commenters and posters. As seen in Figure 7, authentic news posters and commenters share nearly the exact same distribution.

Altogether, we thus observe (especially in contrast to authentic news submissions), that a politically different set of users post misinformation news compared to those that comment on it. While we did not observe that the polarization levels of users who comment on misinformation are substantially different from commenters on authentic users, we did observe that they *are* different from posters of misinformation content.

4.4 Intersection of Misinformation, News Media, Toxicity, and Political Polarization across Subreddits

Finally, having seen the distribution of the political polarization and toxicity among different users and in different environments, we now look if these different characteristics correlate with the amount of misinformation and authentic news in each subreddit. Namely, we seek to determine more broadly if levels of misinformation are correlated with increased levels of toxicity. To do this we rely on our list of *misinfo-oriented* and *mainstream-oriented* websites.

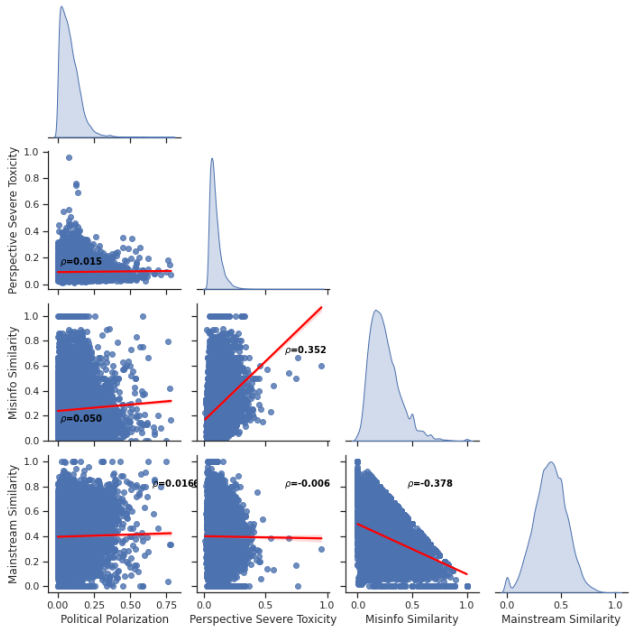


Fig. 8. **Misinformation, toxicity, and political polarization interactions**— As subreddits increase in misinformation levels, they become more toxic. However, there is not a large correlation between misinformation levels and the political polarization of subreddits. Similarly, we do not see any correlation between political polarization levels and mainstream similarity; nor do we see any correlation with toxicity levels.

Specifically, for each subreddit in our dataset, we compute their *misinformation similarity* and their *mainstream similarity* based on the percentage of each subreddit's URL submissions that come from websites that are *misinfo-oriented* and *mainstream-oriented*. This measurement essentially determines the rough approximate percentage of submissions within each of our subreddits that are misinformation oriented/related and the percentage that are mainstream oriented/related.

As seen in Figure 8, across all our 46K considered subreddits, we observe that as subreddits become more similar to misinformation and hyperlink to more *misinfo-oriented* domains, their toxicity increases. This largely matches our observation in Section 4.2 that misinformation submissions are in general more toxic/uncivil than authentic news submissions. Misinformation levels in general thus appear to be correlated with increased toxicity. We further again surprisingly see in Figure 8 that levels of misinformation are not heavily correlated with political polarization. It does not appear that the most politically polarized environments necessarily rely upon misinformation. For example, the most left-leaning subreddits (Table 1) that we observed mostly supported US Senator Bernie Sanders and did not necessarily have high misinformation levels. Conversely, we do not see much of a correlation between subreddits with high mainstream similarity and political polarization and toxicity. This again reinforces our results finding that mainstream news does not have higher levels of toxicity and political polarization level from Section 4.2 and Section 4.3. We thus see again from this analysis that misinformation is indeed correlated with higher toxicity, while authentic news is largely not.

4.5 Summary

In this section, we found that misinformation on Reddit largely is correlative with and predictive of higher amounts of incivility and toxicity on the platform. Most markedly, we observed that the comments under misinformation submissions are posted at a rate 60% higher than the comments under authentic news submissions. Further, while we do observe a dichotomy in the political polarization of users that post misinformation and those that comment on misinformation, somewhat

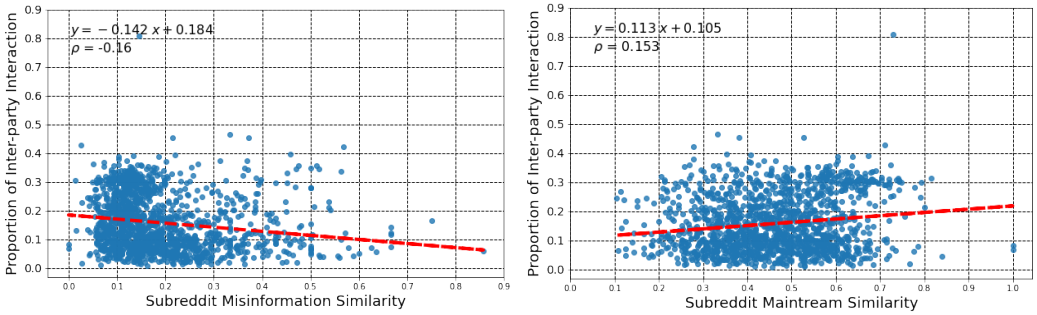


Fig. 9. **Proportions of inter-party interactions in different subreddits**—As subreddits hyperlink to more *misinfo-oriented* websites, as a percentage, there are fewer and fewer interactions between conservative and liberal users. In contrast, there is a slight correlation between hyperlinking to *mainstream-oriented* websites and more inter-party interactions.

surprisingly, we find that misinformation appears across different political environments, with it not being concentrated just in the political extremes. Lastly, looking at how different levels of misinformation correlate with toxicity, we find the more *misinfo-oriented* submissions a given subreddit has, the more toxic/uncivil it is likely to be.

5 RQ2: MISINFORMATION AND POLARIZED TOXIC CONVERSATIONS

As seen in the previous section, comments are 60% more toxic under misinformation submissions than authentic news submissions. Furthermore, there appears to be a difference in the political orientation of users that post misinformation and those that comment on it. Given this difference and the higher toxicity levels present within misinformation submission comments, we now turn to understand if and how these *political differences* drive toxicity within Reddit misinformation submissions.

5.1 Experimental Setup

To fully understand how different levels of *political differences* fuel toxicity and incivility, for this section, we reconstruct the conversational dyads that exist underneath each Reddit submission using the data provided by Pushshift [15]. Comments underneath Reddit submissions are similar to conversational threads; if a user responds to a given comment, their reply will appear underneath the comment. For each submission in our dataset, we thus determine using the thread information whether the commenter posted a response directly to another commenter. This then enables us to reconstruct conversational dyads between individual Reddit users. Then, using the approach outlined in Section 3.1, we determine the polarization and average toxicity of the users in our conversational dyads. From these calculations, we further label users as conservative/conservative-leaning (positive polarization score) or liberal/liberal-leaning (negative polarization based). Lastly, looking at each of these conversational dyads, we determine if each user's response to each other is toxic/uncivil by utilizing the Perspective API SEVERE_TOXICITY classifier with a cutoff of 0.8 (as also outlined in Section 3.1). For a comparison of how conversations differ between misinformation and authentic news comments, we finally separate out the set of conversational dyads that appear under misinformation and authentic news submissions (for both sets of URLs).

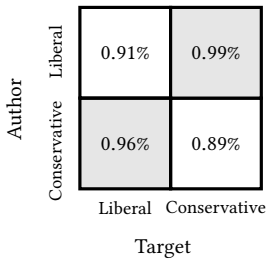


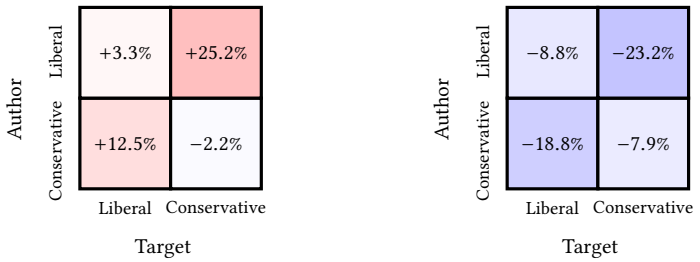
Fig. 10. **Percentage of interactions that are toxic/uncivil for authors and targets of different political leanings.** Across all 46K considered subreddits, there is a slight heterophily for users to reply in a toxic/uncivil manner to members with a tilt towards the opposite political party.

5.2 Toxic Interactions within Misinformation and Authentic News Environments

We find high amounts of homophily in interactions across our conversational dyads. Across all our conversational dyads, indeed 81.3% of interactions are between users of the same political orientation (*i.e* liberal-liberal, conservative-conservative). In contrast, for conversations under mainstream, and misinformation submissions, this increases to 81.7% and 85.3% respectively. We thus see slightly more conversational homophily within misinformation conversations than the entire Reddit population at large. Indeed using our set of *misinfo-oriented* and *mainstream-oriented* and looking at subreddits with at least 500 conversational dyads, as seen in Figure 9, as website hyperlink to more *misinfo-oriented* websites, conversations on their subreddits become more insular (ρ = 0.160). In contrast, as subreddits hyperlink to more *mainstream-oriented* websites, there is a slight increase in the amount of inter-party conversations (ρ = 0.153). We thus see that misinformation is correlated with heightened political homophily within the subreddits, creating more insular communities, while authentic news is associated with a slight increase in political heterophily.

Despite the increased political homophily within misinformation filled-subreddits, we observe a reverse trend in terms of toxic/uncivil comments. As seen in Figure 10, across all considered conversations, we see a slight heterophily for users to reply in a toxic/uncivil manner to users who are not of the same political leaning. We calculate an odds ratio of 1.17 for users to reply in a toxic manner to users of a different political leaning compared with users of the same political leaning. Comparing the set of toxic conversational dyads under misinformation submissions, we see an even higher heterophilic tendency. Compared with the baseline across all conversations, we observe a 25.2% relative increase in the percentage of liberal to conservative toxic comments and a 12.5% relative increase in the percentage of conservative to liberal toxic comments. In contrast, for authentic news submissions, we see a 23.2% relative decrease in liberal to conservative toxic comments and an 18.8% drop in the percentage of conservative to liberal toxic comments. This appears to indicate that while in misinformation-laced conversations, users are more likely to respond in a toxic manner to users of a different political orientation, users in authentic news-centered conversations are less likely.

To confirm, calculating the odds ratio we get 1.64 for misinformation toxic comments and 0.87 for mainstream toxic comments when comparing the percentages of politically heterophilic toxic comments to politically homophilic comments. Overall we thus find that on average, within misinformation submission comments that users are slightly more likely to respond with a toxic comment to users of different political leaning compared to all submissions on Reddit. This largely explains the higher levels of toxicity observed within misinformation submissions in Section 4.2 given the political differences we further observed between misinformation posters and misinformation comments.



(a) Misinformation Submission comments (b) Authentic News Submission comments

Fig. 11. Percentage increases of interactions that are toxic/uncivil in misinformation submissions for conservative and liberal authors against conservative and liberal targets.

5.3 Modeling toxic interactions between users commenting under Misinformation Submissions

In order to confirm the finding that users of different political stripes in misinformation-laced conversations are more likely to reply in a toxic manner to each other, we fit our network data of toxic interactions to an exponential random graph model (ERGM). Exponential Random Graph Models (ERGM) is a form of modeling that predicts connections (toxic interactions) between different nodes (users) in a given network [66]. ERGM models assume that connections are determined by a random variable p^* that is dependent on input variables. As in Chen *et al.* [23] and Peng *et al.* [85], we utilize this modeling as it does it does not assume that its data input is independent; given that we want to model the interactions of polarization, toxicity, this relaxed restriction is key (we have already seen that they are largely not independent) [66, 109]. Utilizing this framework, we thus model the probability of toxic interactions between a given author and target within misinformation submissions as a function of 1) their percentage of toxic comments, 2) their political polarization, 3) the difference in the author and target's political polarization 4) reciprocity between the author and target (*i.e.* if the author and target both had a toxic comment aimed at each other), and finally 5) the number of subreddits that they share.

Toxic Misinformation Submission Interactions	
	Coefficeint
Intercept	***-8.530
Absolute User Polarization	-0.464
User Polarization Differences	***-0.782
User Toxicity	***12.396
Reciprocity	***4.568
Shared Subreddits	0.00051

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 3. As confirmed in our ERGM, differences in political orientation of users is predictive of increased incivility and toxicity. Similarly, the higher each individual user's toxicity norm they are more likely to target other users with toxic comments.

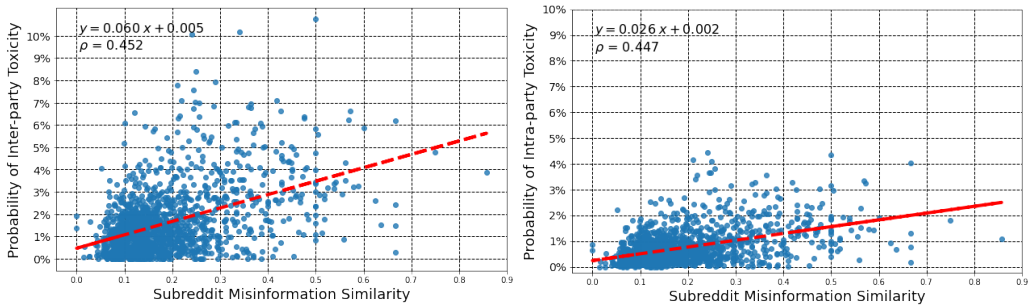


Fig. 12. **Subreddit misinformation similarity vs. probability of toxic interactions between users of different and same political orientation**— While for both inter-political and intra-political interactions, as misinformation similarity in a subreddit increases, the probability of a toxic interaction increases, for inter-political interactions the rate of increase is nearly double.

Fitting our ERGM, we indeed find that as users become more politically distinct from each other, the more likely they are to target each other with uncivil or toxic comments (negative coefficient implies heterophily). As seen in Table 3, political polarization by itself does explain toxicity as found by our model; rather differences between users seem to promote toxicity. This largely reinforces are findings from Section 4. Similarly, as expected, we find that higher levels of user toxicity norms and reciprocity between users are predictive of an increased probability of engaging in toxic interactions. We do not find as users share more subreddits that they are more likely to engage in toxic interactions with each other. We thus have seen that not only do misinformation submissions have more insular conversations, with 85.3% of conversational dyads between users of the same political orientation (compared to 81.3% of conversations under all Reddit submissions) but also that users become more hostile to users of the opposing political orientation.

5.4 Misinformation and increased rates of inter-political toxicity

Finally, having confirmed that users posting under misinformation submission of different political orientations are more likely to engage in negative interactions with each other, we finally determine if the levels of misinformation within given subreddits as a whole leads to increased heterophilic toxic interactions. Namely as misinformation levels in a subreddit as a whole increase does the probability of negative interactions between users of different political orientations increase. We thus now plot the percentage of misinformation within a given subreddit against the probability of toxic interaction between members of the two political orientations.

As seen in Figure 12 looking at subreddits with more than 500 conversational dyads, we see that as subreddits have more *misinformation-oriented* hyperlink submissions, the percentage of conversations dyads between users of different political leanings increases. Concretely, subreddit *misinformation similarity* and the probability with which inter-political conversations are toxic have a correlation of $\rho = 0.452$. While we similarly see that intra-political toxicity also increases with a similar correlation $\rho = 0.447$, we see the rate at which misinformation induces inter-political toxicity is nearly 2.3 times that of intra-political toxicity (0.060 slope vs 0.026 slope). This reflects the fact that *misinformation oriented* submissions are on the whole more toxic but that they increase politically heterophilic toxicity more than politically homophilic toxicity.

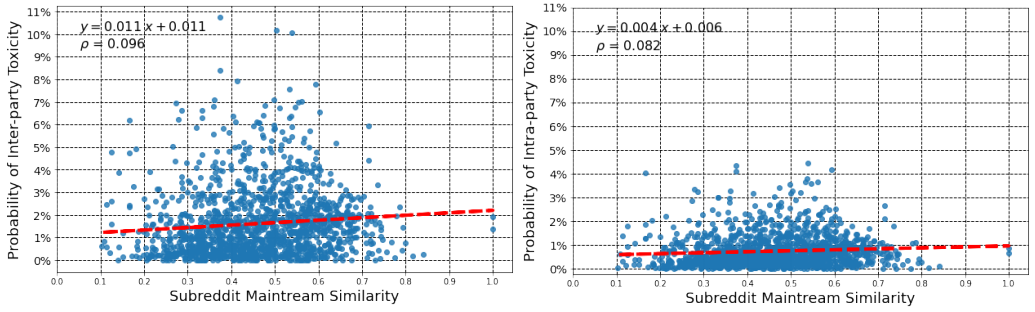


Fig. 13. **Subreddit misinformation similarity vs. probability of toxic interactions between users of different and same political orientation**— For both inter-political and intra-political interactions, as mainstream similarity in a subreddit increases the probability of inter- and intra-political toxicity is largely flat.

Again comparing against subreddit mainstream similarity, we do not see a similar relationship. As seen in Figure 13, the relationship between inter-political and intra-political toxicity rates and similarity to mainstream sources is largely flat.

5.5 Summary

Misinformation, we find, not only promotes higher levels of toxicity in general but also appears to drive inter-political incivility. Fitting an ERGM to our misinformation toxic dyads, we indeed find that political differences (along with reciprocity and each user's toxicity) are a driving force behind the formation of a toxic conversational thread. Finally, examining how different levels of misinformation promote toxicity among users, we find that across our considered subreddits, misinformation drives inter-political incivility at 2.3 times the rate of intra-political toxicity.

6 RQ3: ENGAGEMENT WITH MISINFORMATION AND AUTHENTIC NEWS

Having explored how misinformation on Reddit promotes toxic, insular, and polarized environments on Reddit, we finally turn to understand these factors' role in user engagement with misinformation and authentic news. Namely having seen that misinformation produces more toxic and politically uncivil environments, do these environments lead to more engagement with misinformation? Various works have found that toxic and polarized environments often provoke engagement from users as they get "outraged" by the presented content [45, 71]. In this final section, we seek to determine how different communities and different community norms affect the rates at which misinformation and authentic news get interactions on Reddit.

6.1 Experimental Setup

To understand user interactions and engagement with misinformation and authentic news URL submissions, we utilize the number of comments that each submission receives. We utilize the number of comments rather than the number of upvotes/downvotes, due to the unreliability of Pushshift's data for this particular submission characteristic. While Pushshift often can acquire most submissions and comments, it often fails to keep up-to-date information about the number of votes a given submission receives [15]. This is largely due to the high rate at which submission upvotes/downvotes change. We thus use the more stable and reliable "number of comments" number to determine user engagement with a given submission. We lastly note that to properly model

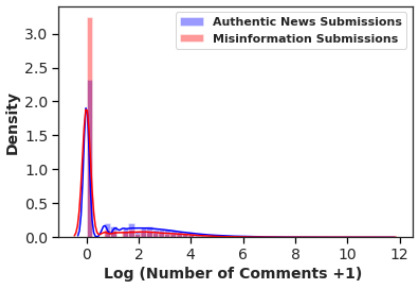


Fig. 14. **Log of number of submissions for misinformation and authentic news submissions**— A large majority of submissions do not receive comments.

the number of comments, we remove comments from Reddit "auto moderator" accounts (often subreddits have auto moderators that automatically comment on submissions). We thus consider the number of comments from "real world" users.

To model the count data of the number of comments on given submissions, we utilize a zero-inflated negative binomial regression [92]. Within our regression, each observation represents a single submission and the number of comments it garnered. We specifically utilize a zero-inflated negative binomial regression as it appropriately models our set of count data. Unlike a Poisson model, which is often utilized to model count data, negative binomial regressions do not make the strong assumption that the mean of the data is equal to the variance [80]. Given that some submissions garner thousands of comments while others garner none, utilizing a Poisson model would be somewhat inappropriate. We further utilize the zero-inflated version of this regression given the heavy preponderance of submissions that do not receive any comments. After removing comments from auto moderators, as depicted in Figure 14, 54.5% of submissions within our dataset did not receive any comments. A normal negative binomial model would thus be unable the correctly model this behavior.

Zero-inflated negative binomial regressions return two sets of coefficients. One set of coefficients, the zero-inflated coefficients, estimated using logistic regression, give the probability that the given submission would receive 0 comments as a function of the covariates. Positive coefficients for these zero-inflated coefficients indicate that increases in the predictor variable make a receiving 0

Number of Comments on Misinfo Submissions		
	Zero Inflated negative coefficient = more likely to get comments	Negative Binomial positive coefficient = more comments
Intercept	5.3146***	3.290***
Absolute User Polarization	-7.2400***	-4.5838***
Absolute Subreddit Polarization	5.3818	0.9658**
Subreddit Toxicity	-2.5780***	5.3042***
User Toxicity	-11.3991 ***	-11.8022***
Average Subreddit Comments	-7.0379 ***	1.0354***
	$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$	

Table 4. Fit of our zero-inflated negative binomial regression on the number of comments on our set of misinformation URL submissions across different subreddits.

comment more likely. Thus the more negative a coefficient, the more the given covariate correlates with inducing at least 1 comment. The second set of coefficients, the negative binomial coefficients, model the number of comments as a function of the covariates. For these coefficients, positive coefficients indicate that the larger the corresponding covariate, the more comments that submission was likely to have received. We thus, in our analysis, can understand how different covariates affect the probability that a given submission will receive *any* comments *and* how these same covariates affect the number of comments received.

For data, we model the number of garnered comments for both our set of 47,822 misinformation submissions and 787,603 authentic news submissions. As factors influencing the number of comments, we utilize (1) the submitter's polarization, (2) the subreddit's polarization, (3) the toxicity norm of the subreddit, (4) the submitter's toxicity norm, and (5) the average number of comments with the subreddit the submission was posted in.

6.2 Results

Before engaging in a thorough analysis of the fits of our zero-inflated negative binomials, we first perform a spot-check on the results. We ensure that the higher the average amount of comments in a given subreddit the more likely a submission is to get comments and that this average correlates with more comments on given submissions. In other words, we check that submissions in subreddits where users comment more, also received more comments. As seen in both Tables 4 and 5, for both misinformation and authentic news Reddit submissions as the average number of comments in a particular subreddit increases, (1) the more likely a submission is to get comments at all and (2) the more comments it is likely to get. Having observed this behavior, we now move to examine the rest of the covariates within our fits.

User Polarization. We observe similar behavior for the effect of the posting user political polarization on the number of comments that the submissions received. For both authentic news and misinformation submissions, we see that as the submission's submitter becomes more politically polarized (*i.e.* moves to the political extremes on the political left or right), the more likely their posts are to receive comments. With zero-inflated coefficients of -7.24 for misinformation submissions

Number of Comments on Authentic News Submissions		
	Zero Inflated negative coefficient = more likely to get comments	Negative Binomial positive coefficient = more comments
Intercept	3.7182***	2.7577***
Absolute User Polarization	-2.7390***	-3.0487 ***
Absolute Subreddit Polarization	5.9227***	1.4898 ***
Subreddit Toxicity	6.1866***	13.1317***
User Toxicity	-14.8969***	−8.0695***
Average Subreddit Comments	-6.4304***	0.6121***

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 5. Fit of our zero-inflated negative binomial regression on the number of comments on our set of authentic news URL submissions across different subreddits.

and -2.74 for authentic news submissions, we see that this is particularly true for misinformation submissions. This largely agrees with prior work that has shown that highly polarized users are likely to provoke and garner comments on social media platforms [62, 71]

However, despite highly polarized users being able to attract at least one comment, we further observe that for both authentic news and misinformation submissions as the posting user becomes more politically polarized, the fewer comments their post is likely to receive. This appears to indicate that in the case of misinformation and authentic news submission, Reddit users are perhaps being "turned off" and engaging less with highly polarized users [60]compared to more politically neutral users.

Subreddit Polarization. While we are unable to conclude using our model whether subreddit polarization has an effect on misinformation submissions, we find that for authentic news submissions the more politically polarized a subreddit is, the less likely anyone is to comment on authentic news submissions. This may indicate as a whole that authentic news submissions do not ordinarily get much traction on highly polarized subreddits. Rather, as documented by Wang *et al.* [114] subreddits like these often ignore more trustworthy sources. We thus find that the insular nature of more partisan subreddits may be inducing them to largely ignore authentic news submissions when they appear within their subreddit.

In contrast, for both misinformation and authentic news submissions, we find that as polarization goes up, the more comments given submissions are likely to garner. This reflects that *when* authentic news and misinformation submissions are noticed, the more polarized the environment, the more users seem to comment and engage with submissions [71].

Subreddit Toxicity. Looking at the subreddit toxicity, we see a marked difference between authentic news submissions and misinformation submissions. We see, notably, for misinformation submissions, the more toxic a subreddit is, the more likely the submission is to get comments. In contrast, for authentic news submissions, the more toxic the subreddit, the more likely the submission is to not get any comments at all. This appears to reflect that authentic news submissions may often not have "clickbait" titles that induce readers to often angrily engage with material [22, 87]. In contrast, oftentimes misinformation websites often post inflammatory articles designed to engender angst in their readership. For example, with regards to the COVID-19 pandemic, the misinformation website infowars.com [102] recently published a report entitled *"Wake Up! Even The Masks Made You Sick"!* As subreddits get more toxic, we thus see that their users are more likely to engage with articles such as this.

Similar to polarization, for both misinformation and authentic news submissions, we find that as subreddit toxicity goes up, the more comments given submissions are likely to garner. This again reflects that *when* authentic news and misinformation submissions are noticed, the more toxic the environment the more users seem to comment and engage with submissions. We thus see that though authentic news submissions are more often ignored in toxic subreddits when compared to misinformation, when they are noticed, toxic environments produce more engagement with both types of submission.

User Toxicity. Finally, looking at user toxicity, we see similar behaviors for both authentic news and misinformation submissions. Most notably, as users become more toxic for both misinformation and authentic news submission, the more likely they are to provoke at least one comment. We thus see that user toxicity is a means by which to gain engagement from posting news articles generally. However, again in both cases, we see that while user toxicity often provokes at least one person to react, we see that this toxicity, often does not lead to more comments on the whole. As found in prior work, toxic users, while often sparking retorts as other users become enraged, also create

unhealthy, short, and otherwise bad conversational outcomes [73, 97]. This result largely matches our definition of individual toxicity as comments that *are likely to make one leave the discussion* from Section 2.1. We thus as a whole see that as entire communities/subreddit becomes more toxic, they are more likely to engage with materials more thoroughly, but as individual users become more toxic, they are more likely to end and otherwise shorten conversations.

7 LIMITATIONS

In this work, we took a large-scale approach to understand the role of misinformation in reinforcing insular and toxic communities online. Given that our approach examines multiple communities simultaneously, there are some tradeoffs we make. We outline and go over some of these limitations here.

One of the limitations of our approach is our use of hyperlinks in order to determine the presence of misinformation and to estimate political polarization levels. Our approach relies on the presence of particular US-based domains on given subreddits and largely only measures US-centric misinformation and polarization. As a result, we are largely unable to extrapolate our results to non-English subreddits and non-US-based political environments. Furthermore, because we examined much of Reddit using our approach we were unable to take a comment-by-comment-based approach to understand the levels of misinformation. As a result our approach inevitably missed out on some of the subtleties of the misinformation in different subreddits. However, we note that as found in several past works [57, 61, 98, 113], examining misinformation from a domain-based perspective enables researchers to track readily-identifiable questionable information across different platforms and thus is a reliable way of understanding the presence of misinformation. Similarly, while our work centers on US-based political environments, as found in prior works, highly political environments across different cultures often utilize misinformation and often share many of the same characteristics as US ones [57, 67].

Another limitation of our approach, given our use of hyperlinks to estimate political polarization and the Perspective API to estimate toxicity, is that it is largely limited to relatively more active users and subreddits. We are only able to develop, in line with past works, toxicity norms and political estimations for subreddits that have at least 50 comments and more than 10 URL submission posts that were in Robertson *et al.*'s dataset. As a result, our results are skewed to subreddits and users that post more often. However, we argue that these subreddits and users make up a large percentage of users' experiences on the Reddit platform and thus accurately model how users interact with each other more generally. For example, our set considered subreddits have interactions from over 59.2% of all active users (posted at least once in the 18-month time frame) on the platform and nearly all of the Reddit comments and submissions. Furthermore, we further note that much of our work was scoped to consider only interactions under authentic news and misinformation submission URLs.

8 DISCUSSION

In this work, we examined how misinformation correlates and is found within more politically insular and toxic environments. Using two lists of misinformation and authentic news domains, we find that the comments underneath Reddit submissions using misinformation websites produced toxic comments 63% more often. Examining how political polarization informs the increase in toxicity within these subreddits, we find, confirming with an ERGM, that misinformation drives toxicity between users of different political leanings. Finally, utilizing a zero-inflated negative binomial model to model engagement with misinformation versus authentic news, we observe that subreddit toxicity is a major predictor of whether misinformation submissions receive comments.

This is in contrast to authentic news submissions which are oftentimes ignored within more politically polarized and toxic/uncivil subreddits.

8.1 Misinformation Enabling Toxicity

As found by Cinelli *et al.* [27], users that post under YouTube videos promoting COVID-19 conspiracy theories, often utilize toxic and vulgar language. Our work has managed to extend their own, illustrating not only that misinformation correlates with increased incivility but also that this increased toxicity often lies in conversations between users of different political orientations conversations. Furthermore, as found with our zero-inflated negative binomial model, subreddit toxicity norms are also predictive of user engagement with misinformation. Misinformation, it appears, promotes and is found within toxic environments. The more toxic/uncivil a given environment the more people appear to engage with misinformation or unreliable sources. The community guidelines for particular environments seem to heavily affect how users engage with different material. We note, furthermore, as found by Gallacher *et al.* [45], toxic online interactions between political groups often lead to offline real-world political violence. Given that misinformation appears to be correlated with and reinforces toxic interactions between different political groups, this highlights the need to research more of its effects and curtail its spread.

8.2 Echo chambers

Much debate has gone into whether online environments and social media platforms are actually "echo chambers." Despite Reddit specifically creating communities for like-minded people and that most interactions on Reddit are amongst people of the same political orientation, in many subreddits we find that a significant proportion of conversations are between users of different political orientations (Figure 9). Further, as users engage more with reliable sources, they are slightly more likely to engage with others across political differences. Furthermore, as found by De Francisci Morales *et al.* [32], Reddit users often engage with each other on more politically neutral subreddits. However, despite these findings, we find, in line with other works [10], that across Reddit that users are more likely to engage in toxic interactions with users of different political orientations. Similarly, as communities become more insular and filled with misinformation, the rate at which users engage in toxic interactions with users of the different political party increases dramatically. We thus find evidence for at least some form of toxic echo chambers, especially among misinformation-filled subreddits.

9 CONCLUSION

We have seen that misinformation persists across many different types of subreddits. Its spread furthermore seems to be affected by the type of community it is posted in. Misinformation appears to be more likely to gain traction and engagement when it is posted in more toxic/uncivil environments. Furthermore, the communities with large amounts of misinformation appear to be more politically insular with most of their interactions occurring between users of similar political orientation. As users become more dissimilar within these misinformation-filled subreddits, as found with our ERGM, they are more likely to be toxic/uncivil to one another. Comparatively, subreddits with less misinformation and more authentic news, are more likely to produce less toxic/uncivil conversations between different types of political users.

Our work, one of the first to examine the relationship between misinformation, toxicity, and political polarization across such a large corpus and in multiple communities, illustrates the need to fully understand the full effect of misinformation. Not only does misinformation and unreliable source sometimes mislead people but they also can magnify political differences and lead to toxic online environments.

REFERENCES

[1] 2021. Twitter. Rules enforcement. https://transparency.twitter.com/en/reports/rules-enforcement.html-2020-jul-dec.

[2] 2022. Google Jigsaw. Perspective API. https://www.perspectiveapi.com/#/home.

[3] 2022. Metrics For Reddit - Complete List Of Subreddits - Updated Weekly. https://frontpagemetrics.com/list-all-subreddits

[4] Sara Abdali, Rutuja Gurav, Siddharth Menon, Daniel Fonseca, Negin Entezari, Neil Shah, and Evangelos E Papalexakis. 2021. Identifying Misinformation from Website Screenshots. In *International AAAI Conference on Web and Social Media (ICWSM) 2021*.

[5] Wasim Ahmed, Josep Vidal-Alaball, Joseph Downing, Francesc López Seguí, et al. 2020. COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data. *Journal of medical internet research* 22, 5 (2020), e19458.

[6] Alexa Internet, Inc. 2021. Top 1,000,000 Sites. http://s3.amazonaws.com/alexa-static/top-1m.csv.zip.

[7] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.

[8] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics* 6, 2 (2019), 2053168019848554.

[9] Jisun An, Daniele Quercia, and Jon Crowcroft. 2014. Partisan sharing: Facebook evidence and societal consequences. In *Proceedings of the second ACM conference on Online social networks*. 13–24.

[10] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[11] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3528–3539.

[12] Pablo Barberá. 2014. How social media reduces mass political polarization. Evidence from Germany, Spain, and the US. *Job Market Paper, New York University* 46 (2014), 1–46.

[13] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* 26, 10 (2015), 1531–1542.

[14] Michael Barthel, Galen Stocking, Jesse Holcomb, and Amy Mitchell. 2016. Reddit news users more likely to be male, young and digital in their news preferences | Pew Research Center. https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

[15] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 830–839.

[16] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Michelangelo Puliga, Antonio Scala, Guido Caldarelli, Brian Uzzi, and Walter Quattrociocchi. 2016. Users polarization on Facebook and Youtube. *PloS one* 11, 8 (2016), e0159641.

[17] Porismita Borah. 2013. Interactions of news frames and incivility in the political blogosphere: Examining perceptual outcomes. *Political Communication* 30, 3 (2013), 456–473.

[18] Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. 2022. Finding Qs: Profiling QAnon Supporters on Parler. https://doi.org/10.48550/ARXIV.2205.08834

[19] Michael A Cacciatore, Dietram A Scheufele, and Shanto Iyengar. 2016. The end of framing as we know it… and the future of media effects. *Mass communication and society* 19, 1 (2016), 7–23.

[20] Pew Research Center. 2017. The partisan divide on political values grows even wider. *Pew Research Center* (2017).

[21] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.

[22] Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*. 15–19.

[23] Yingying Chen and Luping Wang. 2022. Misleading political advertising fuels incivility online: A social network analysis of 2020 US presidential election campaign video comments on YouTube. *Computers in Human Behavior* 131 (2022), 107202.

[24] Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2021. Causal understanding of fake news dissemination on social media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 148–157.

[25] Yun Yu Chong and Haewoon Kwak. 2022. Understanding Toxicity Triggers on Reddit in the Context of Singapore. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1383–1387.

[26] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2020. Echo chambers on social media: A comparative analysis. *arXiv preprint arXiv:2004.09603* (2020).

[27] Matteo Cinelli, Andraž Pelicon, Igor Mozetič, Walter Quattrociocchi, Petra Kralj Novak, and Fabiana Zollo. 2021. Dynamics of online hate and misinformation. *Scientific reports* 11, 1 (2021), 1–12.

[28] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 social media infodemic. *Scientific reports* 10, 1 (2020), 1–10.

[29] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011. Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 192–199.

[30] Dana Cuomo and Natalie Dolci. 2019. Gender-Based Violence and Technology-Enabled Coercive Control in Seattle: Challenges & Opportunities.

[31] Alina Darmstadt, Mick Prinz, and Oliver Saal. 2019. The murder of Keira: misinformation and hate speech as far-right online strategies. (2019).

[32] Gianmarco De Francisci Morales, Corrado Monti, and Michele Starnini. 2021. No echo in the chambers of political interactions on Reddit. *Scientific reports* 11, 1 (2021), 1–12.

[33] Shiri Dori-Hacohen, Keen Sung, Jengyu Chou, and Julian Lustig-Gonzalez. 2021. Restoring Healthy Online Discourse by Detecting and Reducing Controversy, Misinformation, and Toxicity Online. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2627–2628.

[34] Maeve Duggan. 2017. Online Harassment 2017 | Pew Research Center. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

[35] Régis Ebeling, Carlos Abel Córdova Sáenz, Jéferson Campos Nobre, and Karin Becker. 2022. Analysis of the influence of political polarization in the vaccination stance: the Brazilian COVID-19 scenario. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 159–170.

[36] Facebook. 2021. Transparency center. https://transparency.fb.com/policies/community-standards/bullying-harassment/datz. Accessed: 2021-10-08.

[37] Alan Feuer. 2022. In Ad, Shotgun-Toting Greitens Asks Voters to Go 'RINO Hunting' - The New York Times. https://www.nytimes.com/2022/06/20/us/politics/eric-greitens-rino-ad.html

[38] Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.

[39] Christina Fink. 2018. Dangerous speech, anti-Muslim violence, and Facebook in Myanmar. *Journal of International Affairs* 71, 1.5 (2018), 43–52.

[40] Amos Fong, Jon Roozenbeek, Danielle Goldwert, Steven Rathje, and Sander van der Linden. 2021. The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on Twitter. *Group Processes & Intergroup Relations* 24, 4 (2021), 606–623.

[41] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

[42] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2018. "A Stalker's Paradise" How Intimate Partner Abusers Exploit Technology. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[43] Diana Freed, Jackeline Palmer, Diana Elizabeth Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2017. Digital technologies and intimate partner violence: A qualitative analysis with multiple stakeholders. *Proceedings of the ACM on human-computer interaction* 1, CSCW (2017), 1–22.

[44] Daniel Funke. 2018. Fact-checkers have debunked this fake news site 80 times. It's still publishing on Facebook. Poynter. org.

[45] John D Gallacher, Marc W Heerdink, and Miles Hewstone. 2021. Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters. *Social Media+ Society* 7, 1 (2021), 2056305120984445.

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

[47] Anthony J Gaughan. 2016. Illiberal democracy: The toxic mix of fake news, hyperpolarization, and partisan election administration. *Duke J. Const. L. & Pub. Pol'y* 12 (2016), 57.

[48] Bryan T Gervais. 2015. Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics* 12, 2 (2015), 167–185.

[49] Dipayan Ghosh and Ben Scott. 2018. Digital deceit: the technologies behind precision propaganda on the internet. (2018).

[50] Amit Goldenberg and James J Gross. 2020. Digital emotion contagion. *Trends in Cognitive Sciences* 24, 4 (2020), 316–328.

[51] Ine Goovaerts and Sofie Marien. 2020. Uncivil communication and simplistic argumentation: Decreasing political trust, increasing persuasive power? *Political Communication* 37, 6 (2020), 768–788.

[52] Kirsikka Grön and Matti Nelimarkka. 2020. Party Politics, Values and the Design of Social Media Services: Implications of political elites' values and ideologies to mitigating of political polarisation through design. *Proceedings of the ACM on human-computer interaction* 4, CSCW2 (2020), 1–29.
[53] Anatoliy Gruzd and Philip Mai. 2020. Going viral: How a single tweet spawned a COVID-19 conspiracy theory on Twitter. *Big Data & Society* 7, 2 (2020), 2053951720938405.
[54] Andrew Guess, Brendan Nyhan, and Jason Reifler. 2018. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. *European Research Council* 9, 3 (2018), 4.
[55] Yosh Halberstam and Brian Knight. 2016. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of public economics* 143 (2016), 73–88.
[56] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 202. A Golden Age: Conspiracy Theories' Relationship with Misinformation Outlets, News Media, and the Wider Internet. (202).
[57] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 2022. No Calm in The Storm: Investigating QAnon Website Relationships. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 299–310.
[58] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 2023. Happenstance: Utilizing Semantic Search to Track Russian State Media Narratives about the Russo-Ukrainian War On Reddit. 17 (2023).
[59] Gordon Heltzel and Kristin Laurin. 2020. Polarization in America: Two possible futures. *Current Opinion in Behavioral Sciences* 34 (2020), 179–184.
[60] Marc J Hetherington. 2008. Turned off or turned on? How polarization affects political engagement. *Red and blue nation* 2 (2008), 1–33.
[61] Austin Hounsel, Jordan Holland, Ben Kaiser, Kevin Borgolte, Nick Feamster, and Jonathan Mayer. 2020. Identifying Disinformation Websites Using Infrastructure Features. In *USENIX Workshop on Free and Open Communications on the Internet*.
[62] Philip N Howard, Bharath Ganesh, Dimitra Liotsiou, John Kelly, and Camille François. 2019. The IRA, social media and political polarization in the United States, 2012-2018. (2019).
[63] Yiqing Hua, Mor Naaman, and Thomas Ristenpart. 2020. Characterizing twitter users who engage in adversarial interactions against political candidates. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
[64] Y Linlin Huang, Kate Starbird, Mania Orand, Stephanie A Stanek, and Heather T Pedersen. 2015. Connected through crisis: Emotional proximity and the spread of misinformation online. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 969–980.
[65] Robert Huckfeldt, Paul Allen Beck, Russell J Dalton, and Jeffrey Levine. 1995. Political environments, cohesive social groups, and the communication of public opinion. *American Journal of Political Science* (1995), 1025–1054.
[66] David R Hunter, Mark S Handcock, Carter T Butts, Steven M Goodreau, and Martina Morris. 2008. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of statistical software* 24, 3 (2008), nihpa54860.
[67] Roland Imhoff, Felix Zimmer, Olivier Klein, João HC António, Maria Babinska, Adrian Bangerter, Michal Bilewicz, Nebojša Blanuša, Kosta Bovan, Rumena Bužarovska, et al. 2022. Conspiracy mentality and political orientation across 26 countries. *Nature human behaviour* 6, 3 (2022), 392–403.
[68] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. " Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–33.
[69] Shan Jiang and Christo Wilson. 2018. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.
[70] Julia Kamin. 2019. *Social Media and Information Polarization: Amplifying Echoes or Extremes?* Ph. D. Dissertation.
[71] Jin Woo Kim, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication* 71, 6 (2021), 922–946.
[72] Yonghwan Kim and Youngju Kim. 2019. Incivility on Facebook and political polarization: The mediating role of seeking further comments and negative emotion. *Computers in Human Behavior* 99 (2019), 219–227.
[73] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing Toxic Content Classification for a Diversity of Perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. 299–318.
[74] K Hazel Kwon and Anatoliy Gruzd. 2017. Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research* (2017).
[75] Charlotte Lambert, Ananya Rajagopal, and Eshwar Chandrasekharan. 2022. Conversational Resilience: Quantifying and Predicting Conversational Outcomes Following Adverse Events. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 548–559.

[76] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.

[77] Lucas Lima, Julio CS Reis, Philipe Melo, Fabricio Murai, Leandro Araujo, Pantelis Vikatos, and Fabricio Benevenuto. 2018. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 515–522.

[78] Daniela Mahl, Jing Zeng, and Mike S Schäfer. 2021. From "Nasa Lies" to "Reptilian Eyes": Mapping Communication About 10 Conspiracy Theories, Their Communities, and Main Propagators on Twitter. *Social Media+ Society* 7, 2 (2021), 20563051211017482.

[79] Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. 2020. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–24.

[80] Durim Morina and Michael S Bernstein. 2022. A Web-Scale Analysis of the Community Origins of Image Memes. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–25.

[81] Ashley Muddiman, Shannon C McGregor, and Natalie Jomini Stroud. 2019. (Re) claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication* 36, 2 (2019), 214–226.

[82] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. 145–153.

[83] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2020. Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. In *AAAI Conference on Web and Social Media*.

[84] Marius Paraschiv, Nikos Salamanos, Costas Iordanou, Nikolaos Laoutaris, and Michael Sirivianos. 2022. A Unified Graph-Based Approach to Disinformation Detection using Contextual and Semantic Relations. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 747–758.

[85] Tai-Quan Peng, Mengchen Liu, Yingcai Wu, and Shixia Liu. 2016. Follower-followee network, communication networks, and vote agreement of the US members of congress. *Communication research* 43, 7 (2016), 996–1024.

[86] Nathaniel Persily. 2017. The 2016 US Election: Can democracy survive the internet? *Journal of democracy* 28, 2 (2017), 63–76.

[87] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *European conference on information retrieval*. Springer, 810–817.

[88] Walter Quattrociocchi, Rosaria Conte, and Elena Lodi. 2011. Opinions manipulation: Media, power and gossip. *Advances in Complex Systems* 14, 04 (2011), 567–586.

[89] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. 2016. Echo chambers on Facebook. *Available at SSRN 2795110* (2016).

[90] Stephen A Rains, Kate Kenski, Kevin Coe, and Jake Harwood. 2017. Incivility and political identity on the Internet: Intergroup factors as predictors of incivility in discussions of news online. *Journal of Computer-Mediated Communication* 22, 4 (2017), 163–178.

[91] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 557–568.

[92] Martin Ridout, John Hinde, and Clarice GB Demétrio. 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* 57, 1 (2001), 219–223.

[93] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22.

[94] Daniel Romer and Kathleen Hall Jamieson. 2020. Conspiracy theories as barriers to controlling the spread of COVID-19 in the US. *Social science & medicine* 263 (2020), 113356.

[95] Martin Saveski, Doug Beeferman, David McClure, and Deb Roy. 2022. Engaging Politically Diverse Audiences on Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 873–884.

[96] Martin Saveski, Nabeel Gillani, Ann Yuan, Prashanth Vijayaraghavan, and Deb Roy. 2022. Perspective-taking to reduce affective polarization on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 885–895.

[97] Martin Saveski, Brandon Roy, and Deb Roy. 2021. The structure of toxic conversations on Twitter. In *Proceedings of the Web Conference 2021*. 1086–1097.

[98] Vibhor Sehgal, Ankit Peshin, Sadia Afroz, and Hany Farid. 2021. Mutual hyperlinking among misinformation peddlers. *arXiv preprint arXiv:2104.11694* (2021).

[99] Karishma Sharma, Emilio Ferrara, and Yan Liu. 2022. Construction of Large-Scale Misinformation Labeled Datasets from Social Media Discourse using Label Refinement. In *Proceedings of the ACM Web Conference 2022*. 3755–3764.
[100] Karishma Sharma, Yizhou Zhang, and Yan Liu. 2022. COVID-19 Vaccine Misinformation Campaigns and Social Media Narratives. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 920–931.
[101] Cuihua Shen, Qiusi Sun, Taeyoung Kim, Grace Wolff, Rabindra Ratan, and Dmitri Williams. 2020. Viral vitriol: Predictors and contagion of online toxicity in World of Tanks. *Computers in Human Behavior* 108 (2020), 106343.
[102] Kate Starbird, Ahmer Arif, Tom Wilson, Katherine Van Koevering, Katya Yefimova, and Daniel Scarnecchia. 2018. Ecosystem or echo-system? Exploring content sharing across alternative media domains. In *Proceedings of the International AAAI Conference on Web and Social Media*.
[103] Jennifer Stromer-Galley. 2003. Diversity of political conversation on the Internet: Users' perspectives. *Journal of Computer-Mediated Communication* 8, 3 (2003), JCMC836.
[104] Cass R Sunstein. 2018. Is social media good or bad for democracy. *SUR-Int'l J. on Hum Rts.* 27 (2018), 83.
[105] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 247–267.
[106] Christopher Torres-Lugo, Kai-Cheng Yang, and Filippo Menczer. 2022. The Manufacture of Partisan Echo Chambers by Follow Train Abuse on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1017–1028.
[107] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* (2018).
[108] Joshua A Tucker, Yannis Theocharis, Margaret E Roberts, and Pablo Barberá. 2017. From liberation to turmoil: Social media and democracy. *Journal of democracy* 28, 4 (2017), 46–59.
[109] Johannes van der Pol. 2019. Introduction to network modeling using exponential random graph models (ergm): theory and an application using R-project. *Computational Economics* 54, 3 (2019), 845–875.
[110] Chris J Vargo and Toby Hopp. 2017. Socioeconomic status, social capital, and partisan polarity as predictors of political incivility on Twitter: a congressional district-level analysis. *Social Science Computer Review* 35, 1 (2017), 10–32.
[111] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)* 13, 2 (2019), 1–22.
[112] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.
[113] Elliott Waissbluth, Hany Farid, Vibhor Sehgal, Ankit Peshin, and Sadia Afroz. 2022. Domain-Level Detection and Disruption of Disinformation. *arXiv preprint arXiv:2205.03338* (2022).
[114] Yuping Wang, Savvas Zannettou, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, and Gianluca Stringhini. 2021. A Multi-Platform Analysis of Political News Discussion and Sharing on Web Communities. In *IEEE Conference on Big Data*.
[115] Brian E Weeks. 2015. Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of communication* 65, 4 (2015), 699–719.
[116] Galen Weld, Amy X Zhang, and Tim Althoff. 2022. What Makes Online Communities 'Better'? Measuring Values, Consensus, and Conflict across Thousands of Subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1121–1132.
[117] Tom Wilson and Kate Starbird. 2020. Cross-platform disinformation campaigns: Lessons learned and next steps. *Harvard Kennedy School Misinformation Review* (2020).
[118] Magdalena E Wojcieszak and Diana C Mutz. 2009. Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of communication* 59, 1 (2009), 40–56.
[119] Michael J Wood. 2018. Propagating and debunking conspiracy theories on Twitter during the 2015–2016 Zika virus outbreak. *Cyberpsychology, behavior, and social networking* 21, 8 (2018), 485–490.
[120] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*. 1391–1399.
[121] Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. 2020. Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-computer Interaction* 4, CSCW2 (2020), 1–23.
[122] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 internet measurement conference*. 405–417.

[123] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1350–1361.

[124] Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. 2020. Quantifying the causal effects of conversational tendencies. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–24.