Sub-Standards and Mal-Practices: Misinformation's Role in Insular, Polarized, and Toxic Interactions on Reddit

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

The spread of unreliable information online has been linked to rising political hostility, yet its specific relationship to user behavior and discourse remains underexplored. In this work, we examine the influence of unreliable information on political incivility and toxicity on the social media platform Reddit. We show that comments on articles from unreliable news websites are posted more often in right-leaning subreddits and that, within individual subreddits, comments are on average 32% relatively more likely to be toxic compared with comments on reliable news articles. Utilizing a zero-inflated negative binomial regression, we further show that as the toxicity of subreddits increases, users are more likely to comment on posts from known unreliable websites. Finally, modeling user interactions with an exponential random graph model, we find that when reacting to a Reddit submission linking to a website known for spreading unreliable information, users are more likely to be toxic toward users of different political beliefs. Our results collectively illustrate that low-quality, unreliable information not only predicts increased toxicity but also fosters more polarizing interactions between users of different political orientations.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; **Empirical studies in collaborative and social computing**; • **Information systems** → **World Wide Web**.

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

ACM Reference Format:
Hans W. A. Hanley and Zakir Durumeric. 2025. Sub-Standards and Mal-Practices: Misinformation's Role in Insular, Polarized, and Toxic Interactions on Reddit. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW480 (November 2025), 35 pages. https://doi.org/10.1145/3757661

1 Introduction

Content Warning: This paper studies online toxicity. When necessary for clarity, this paper quotes user content that can be considered profane, politically inflammatory, and hateful.

Over the last decade, misinformation, incivility, and political polarization have corroded the public's trust in democratic institutions [17, 27, 54, 55, 58]. Despite their shared roles in disrupting discourse and stoking political division, misinformation, online toxicity, and polarization are separate phenomena, and their complex interactions remain debated and somewhat unclear [22, 31, 34, 38, 57, 109, 142, 152, 157]. For instance, recent work from Quattrociocchi et al. [121] found that on X (formerly Twitter), toxic language is equally distributed across conversations regardless of the presence of reliable or unreliable news. Similarly, Cinelli et al. [31] found that "there are no significant differences between the proportions of hate speech detected in comments on videos

Authors' Contact Information: Hans W. A. Hanley, hhanley@cs.stanford.edu, Stanford University, Stanford, California, USA; Zakir Durumeric, zakir@cs.stanford.edu, Stanford University, Stanford, California, USA.

ACM 2573-0142/2025/11-ARTCSCW480
https://doi.org/10.1145/3757661

from questionable and reliable channels" on YouTube. In contrast, Mosleh et al. [111, 112] found that false headlines on X are correlated with increased toxicity and Dicicco et al. [37] found that throughout the COVID-19 pandemic, conspiracy theories emerged amongst users who regularly employed toxic language.

In this work, we investigate the interplay of toxicity, partisanship, and unreliable information in a structured environment: Reddit. As found in prior work, unlike relatively unstructured platforms like X and YouTube [117, 132], Reddit communities called subreddits have relatively distinct and stable political and toxicity norms [95, 107, 125], allowing for more direct study of the complex interplay of toxicity, partisanship, and unreliable information. Due to the structure of Reddit and as found by An et al. [10], Reddit communities are fairly homogeneous, with users gathering based on shared interests [105] (*i.e.*, the users on r/democrats are largely left-leaning compared to the users on r/Conservative who tend to be right-leaning). Furthermore, as also found by An et al. [10], even in "cross-cutting" communities where users are encouraged to interact with users with different views (*e.g.*, /r/PoliticalDebate), Reddit users rarely engage with their partisan opposites on the same threads [16]. The particular structure of these Reddit communities enables us to understand (1) if the characteristics of stable communities and their constituent users promote toxic interactions and engagement (2) if unreliable information potentially increases the likelihood of toxicity and (3) if, in the rare occasions where users interact with other accounts of different partisan orientations, this interaction promotes toxicity and partisan rancor. By studying these phenomena jointly over a broad period of time across all of Reddit's stable communities, we argue that we can further understand the characteristics of and the influence of unreliable information and political polarization in provoking bad outcomes like toxic behavior on Reddit. Furthermore, unlike past studies [95, 105, 115, 125], our work approaches understanding toxic behavior by simultaneously considering Reddit interactions at a community, thread, and user level. Furthermore, in contrast to prior work, which has been limited to explicitly political settings, we analyze a diverse set of subreddits, measuring the influence of misinformation on toxicity while accounting for the "politicalness" of each community [107]. Concretely, we ask the following research questions:

(1) *RQ1: Do Reddit posts linking to articles from unreliable websites have increased toxicity and engagement? How do subreddit norms (e.g., political partisanship) predict toxic engagement?*
(2) *RQ2: Does unreliable news exacerbate toxic interactions between users with political partisanship differences (i.e., affective polarization)?*

Practically, we propose the workflow visualized in Figure 1. First, we gather all Reddit data from a period of 18 months (January 2020 to June 2021), encompassing a collection of over 1.6 billion comments. We then determine the number of toxic comments within each subreddit and from individual users using the Google Jigsaw Perspective API [86], a commonly deployed classifier for identifying toxic language. Then, utilizing a Word2Vec approach from Waller et al. [160], we approximate the partisanship and the "politicalness" of a subset of subreddits and users along the US left–right political spectrum. Finally, we utilize previously curated lists of reliable and unreliable news sites to determine the levels at which communities and users link to websites known to spread misinformation. From these calculations, we analyze the relationships between toxicity, political partisanship, and misinformation.

RQ1: Toxicity, Partisanship, and Unreliable Information. We first determine whether there are distinct levels of user political partisanship and toxicity in the comments that respond to articles from unreliable versus reliable news outlets. We find that comments posted on articles from unreliable websites are, on average, 32% more toxic within individual subreddits and 25% more toxic across Reddit as a whole than comments responding to reliable websites. Fitting a linear regression with LASSO regularization against the average toxicity of users' comments, we

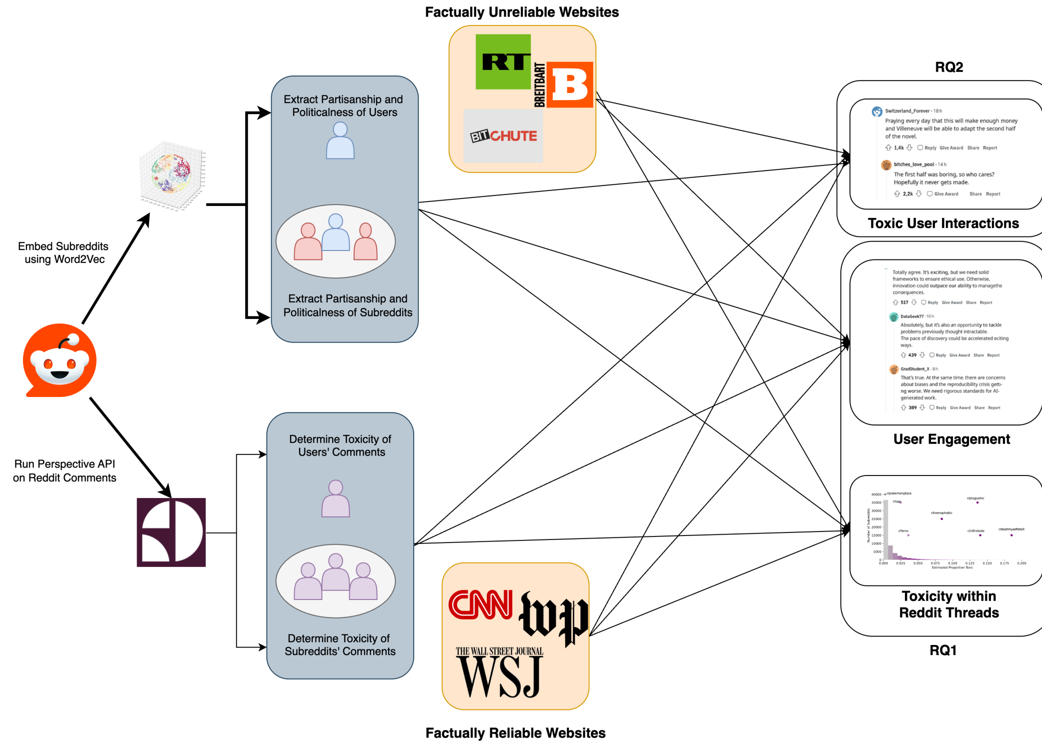


Fig. 1. We measure the partisanship, politicalness, and the prevalence of unreliable and reliable news to understand the interplay of these variables in predicting levels of toxicity within Reddit threads, overall user engagement, and undesirable toxic interactions between users.

find that partisanship, each subreddit/community's toxicity norms, and the involvement of Reddit moderators/admins in a thread predict the toxicity of conversations. Finally, we show that as subreddits become more toxic, users are more likely to comment on unreliable news articles.

RQ2: Engagement with Unreliable News Sources Predicting Inter-Political Strife. Having identified that users who comment on unreliable sources are more likely to post toxic comments than those who respond to reliable website posts, we examine the role of political partisanship in these toxic interactions. We find that users who comment under Reddit submissions to unreliable sources have a higher rate of inter-partisan toxicity compared to users who comment under reliable sources (1.38 odds ratio). Indeed, using an Exponential Random Graph model to understand collective user interactions, we find that users who comment on unreliable domain submissions are more likely to respond to users of different political views in a toxic manner and to reciprocate toxic comments aimed at them.

Altogether, we show partisanship, affective polarization, and unreliable websites' role in promoting toxicity on Reddit and insular online communities. Our work, one of the first to examine the relationship between unreliable news sources, toxicity, and political partisanship within and between different communities of varying levels of political engagement, illustrates the need to and how to fully understand the complex interactions between these phenomena so that platforms can better understand and address toxicity online.

2 Background & Related Work

In this section, we detail key definitions, provide background on Reddit, and give an overview of prior works that analyze the effects of misinformation, toxicity, and political polarization on social media.

2.1 Terminology

Building on extensive prior work on misinformation, toxicity, and political polarization [30, 61, 65, 148], we utilize community-accepted definitions of the following terms:

Reliable and Unreliable Domains. As in previous studies [7, 62, 65, 78, 85, 102, 162], we define *misinformation* as information that is false or inaccurate regardless of author intention. Similarly, we define *unreliable domains* as websites that regularly publish false information about current events and that do not engage in journalistic norms such as attributing authors and correcting errors [3, 7, 28, 65, 75, 116, 140, 171]. Conversely, we define *reliable domains* as websites that generally adhere to journalistic norms including attributing authors and correcting errors; altogether publishing mostly true information [65, 75, 171].

Online Toxicity and Incivility. Given our use of the Google Jigsaw Perspective API [86], we use their definition of toxicity: "*(explicit) rudeness, disrespect or unreasonableness of a comment that is likely to make one leave the discussion.*"

Partisanship. We define partisanship as users' and communities' place on the US left–right political spectrum [130]. We note the limitation of this definition given the variety of political views within the US. However, in line with previous work [67, 134, 135], we utilize this definition, which largely fits much of US-centered political discussion, to understand how right-leaning and left-leaning users and communities interact with one another and different news sources.

Affective Political Polarization. Affective political polarization is the tendency of individuals to distrust and be negative towards those of different political beliefs while being positive towards people of similar political views [39].

2.2 Reddit

Reddit is an online social media platform composed of millions of subcommunities known as subreddits [2, 25]. Subreddits are dedicated to specific topics, ranging from politics (r/politics) and science (r/science) to Pokémon (r/pokemon). Depending on the community, users can submit news articles, opinions, images, and memes as *submissions*. Underneath these submissions, other users can comment or reply to comments from other users. Anyone can create a subreddit, and subreddits are moderated by Reddit's content policies, subreddit-specific rules, and implicit community norms [25, 45, 83]. Subreddit norms vary widely [164] and encompass political behaviors, tolerance to misinformation, and toxic behavior [25, 83, 125, 164].

2.3 Partisanship and Polarization

People, both in real life and on the Internet, tend to associate with like-minded people [12, 13, 63, 66, 79, 89, 122]. Wojcieszak et al. [166] find that while the majority of political discussions online are between participants who share the same viewpoint, many users *do* enjoy conversations with people with different viewpoints [146]. Despite this, past works have found that social media platforms may be one of the drivers of political polarization and "echo chambers" [22, 24, 71, 89, 145]. Sunstein, Garrett et al., and Quattrociocchi et al. all argue that the "individualized" experience offered by social media platforms comes with the risk of creating "information cocoons" and "echo chambers" that accelerate polarization [53, 123, 147]. Conover et al. [32] find that X's structure fosters increased levels of politically polarized conversations. Bessi et al. [15], examining the behaviors of 12 million users, find that partisan echo chambers are driven by the algorithms of both Facebook and YouTube. Finally, Torres et al. [149] further find that the specific X behavior of "follow trains" induces highly politically polarized behavior on the platform.

In a similar vein, prior work has found that the increased political polarization engendered by social media causes several negative downstream effects, including the increased sharing of misinformation and toxic online behaviors. Imhoff et al. [82], for example, find that political polarization is associated with beliefs in conspiracy theories. Ebling et al. [42] similarly find that political partisanship levels on social media are associated with medical misinformation. Furthermore, other studies have interrogated the adverse effects that social media has had on the democratic process due to increased political polarization [60, 119, 152, 153].

2.4 Misinformation

Misinformation has increasingly become a major aspect of the conversations on social media [7, 51, 56]. Even after controlling for cascade size, Juul and Ugander find that false information spreads deeper and wider on X than true information. This false information is often spread by relatively few actors. Weld et al. find that over 99% of "extremely biased and low factual content hyperlinks" were shared in only 0.5% of Reddit communities, creating factually unreliable echo chambers [163]. Furthermore, misinformation often convinces those who are exposed to it. A large percentage of US adults were exposed to misinformation stories through social media during the 2016 election [7], and many believed these false stories [6, 62]. As COVID-19 spread throughout the world, online misinformation and conspiracy theories became a major hurdle to curbing its spread [131, 141].

To prevent the spread of misinformation, recent research has focused on several different approaches for tracking and stemming its flow [65, 152]. For example, Mahl et al. [106] employ a network analysis of co-occurring hashtags to identify and track the spread of 10 conspiracy theories on X. Ahmed et al. [4] use a similar approach to track the spread of COVID-19 and 5G conspiracy theories. They find that well-known misinformation websites were some of the largest sources spreading these conspiracy theories on X. Gruzd [61] found that a single Tweet about "how COVID-19 was a hoax" spurred an entire conspiracy theory, eventually prompting large groups of people to film their local hospitals to prove that COVID-19 was not real. In addition to network-based approaches, others have used advancements in natural language processing to identify and track misinformation. Hanley et al. [68], for example, utilize semantic search to identify and track Russian state-media narratives on Reddit. Fong et al. [47] utilize linguistic and social features to understand the psychology of X users who engaged with known conspiracy theorists. Finally, several works have performed in-depth case studies on the spread of specific false narratives: Wilson and Starbird et al. look at the Syrian White Helmets on X and Bär et al. look at the spread of QAnon on Parler [21, 165].

2.5 Toxicity

Online toxicity takes many forms, including threats, sexual harassment, doxing, coordinated bullying, and political incivility [49, 50, 104, 148]. Toxic comments, in particular, are one of the most common forms of hate and harassment online [148] and are seemingly an inescapable part of social media [33, 96, 114, 148, 167]. Past studies have found that 41% of Americans and 40% of those globally have experienced bullying or harassment online [40, 148]. Facebook estimates that 0.14–0.15% of all views on their platform are of toxic comments [44]. This type of incivility, in addition to damaging online conversations, has been found to also damage civil institutions [17, 153], having dangerous real-world implications. For example, Fink et al. [46] found that politically charged anti-Muslim hate speech on Facebook in Myanmar was a prominent aspect preceding the Rohingya genocide.

To limit toxicity, platforms have designed and implemented a variety of safeguards [1, 44, 86]. Other researchers have further performed in-depth studies on users' behavior to understand abusers and victims of abuse. For instance, Founta et al. [48] identify and study a set of network and account characteristics of abusive accounts on X. Hua et al. [77] look at properties of the accounts on X that have heavily negative interactions with political candidates. 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 [99, 168, 172, 173].

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

Given that online toxicity, for instance, has been heavily associated with increased political polarization and misinformation [31, 152], several works have attempted to understand how political partisanship, online toxicity, and misinformation interact. Rajadesingan et al. [124], find that political discussions in non-overtly political subreddits often lead to less toxic conversations. Cinelli et al. [31], show that misinformation about COVID-19 on YouTube promotes hate and toxicity. Chen et al. [27], utilizing network-based analysis, find that misleading online videos often lead to increased incivility in their comments. De Francisci Morales et al. [36] find, most markedly, that the interaction of individuals of different political orientations increased negative conversational outcomes. Similarly, Kim et al., Kwon et al., and Shen et al. find that exposure to negative conversations increases observers' tendency to further engage in incivility [92, 98, 142]. Finally, Imhoff et al. [82] find that political polarization is a key aspect in people's belief in false narratives. However, despite this panoply of past research, it remains unclear how political partisanship and toxicity interact in the presence of misinformation and across political environments. In this work, we seek to understand this dynamic.

2.7 Present Work

While several previous works have studied partisanship and affective polarization [36, 43, 107], finding evidence of inter-partisan hostility, these works have been limited to explicitly politically-oriented spaces and do not study the influence of unreliable information or misinformation. Limiting the study of how partisanship and unreliable information affect users' discussions to only overly political subreddits, as in past works, can thus give an incomplete picture of user behavior. As shown by Rajadesingan et al. [124] and Mamakos et al. [107], political discussions frequently take place in non-overtly political subreddits. As found by Efstratiou et al., different subreddits can have different "echo chamber-like" behaviors and inter-partisan discussions depending on their politicalness [43].

Our work seeks to understand how partisanship and unreliable news sources that spread largely non-factual information contribute to this toxicity and user engagement in both political and non-political contexts. Given that our work quantifies the politicalness and other characteristics of a subreddit or a user, we can account for this factor in contributing to toxicity and explore how unreliable sources interact in different subreddit environments and across different community standards. By examining how these unreliable and reliable sources differ in toxicity both within and between individual subreddits and across subreddits of different types of politicalness, we seek to holistically understand the extent to which the spread of unreliable news correlates with toxicity and increased engagement.

3 Datasets and Methods

In this section, we provide an overview of our datasets, describe how we calculate the political partisanship and politicalness of users and subreddits, how we determine the toxicity of posts and comments, and how we identify user interactions with unreliable and reliable website sources.

3.1 Reddit Dataset

We study 18 months of Reddit comments and submissions from January 2020 to June 2021, which we collected using Pushshift [14]. We note that all data was collected before Pushshift fell outside Reddit's Terms of Service in April 2023. Altogether, we gathered 2.2 billion comments and 491 million submissions. Each comment and submission includes its timestamp, author's username, subreddit, and the conversation thread where the comment was posted. Using this data, we reconstruct the conversation threads for each user and subreddit.

Filtering and Constructing Final Reddit Dataset. As in Kumar et al. [95], given that many Reddit comments are labeled as toxic by models like the Perspective API [86], models trained on the Civil Comments dataset [64], or large language models [94] are simply sexually explicit and contained within 18+ communities, we exclude 18+ subreddits from our study. As argued by Kumar et al. [95], while toxic behaviors do occur within these subreddits, the explicit allowance of sexually explicit language leads to a large number of false positives, complicating analysis. In addition to filtering out 18+ subreddits, we limit our analysis to English and thus filter our dataset using the `whatlanggo` Go language library[1] to only English-language comments. Given that we use the Perspective API to detect toxicity, we limit our analysis to comments that are 15–300 characters in length [96]. We note that this initial filtering phase resulted in an intermediate dataset of 1.8B comments. Altogether, our final dataset consists of 327M Reddit submissions, 1.6B comments, and 15.5M users from 57.2K subreddits.

3.2 Unreliable and Reliable Domain Dataset

To analyze how users interact with misinformation, we first gather a set of unreliable and reliable websites (as a control). Specifically, we aggregate a list of unreliable/misinformation and reliable/authentic-news domains from Media-Bias/Fact-Check.[2] We note that this dataset is utilized widely within the studies of unreliable information online [19, 69, 70, 113, 163]. We utilize a dichotomy between websites on Media-Bias/Fact-Check that promote mostly highly factual information and websites that produce mostly unreliable information. We utilize this binary, given the lack of large differences between the different categories of Media-Bias/Fact-Check. For example, the "Mostly Factual" category includes websites like cnn.com, washingtonpost.com, and wsj.com, all websites that are considered to be *highly* reputable [41, 67]; this is compared to nytimes.com which has a rating of "High" factuality and yahoo.com (which often is an aggregator of news from cnn.com, washingtonpost.com, and wsj.com [170]) which also has a rating of "High." Using this dichotomy has the benefit of enabling us to understand the broad differences between user and subreddit interactions in the presence of articles from unreliable and reliable sources and to not be overly reliant on the fine-grained distinctions from Media-Bias/Fact-Check. Specifically, we consider websites as "unreliable" if their factfulness rating from Media-Bias/Fact-Check is "Low" or "Very Low"; conversely, we consider a website as "reliable" if its factuality rating from Media-Bias/Fact-Check is "Mostly Factual", "High", or "Very High". Again, we include "Mostly Factual" in this category, given that it includes websites like cnn.com and washingtonpost.com. To ensure consistency, we further cross-reference these two lists of websites against news websites previously gathered by Iffy News,[3] OpenSources,[4] Politifact,[5] Snopes,[6] Melissa Zimdars,[7] and Hanley et al. [67].

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

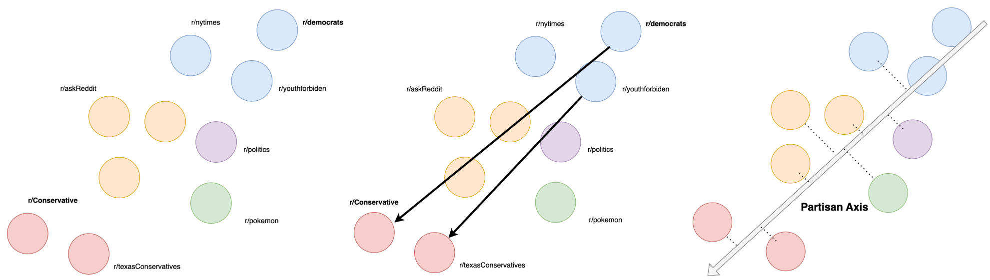


Fig. 2. We determine the partisanship of different subreddits using Waller et al.'s [160] method— Waller et al.'s method entails (1) determining a seed pair of communities that differ primarily in the social dimension of interest, (2) finding pairs of Reddit communities that differ primarily in the same social dimension, and (3) averaging the differences between all the selected pairs as the vector **pr** and projecting each subreddit onto that averaged dimensional vector **pr**.

Our final list of factually unreliable news outlets consists of 1,054 websites, which encompass sites like theconservativetreehouse.com and infowars.com [67]. Separately, our list of reliable news sites consists of 3,754 websites from across the political spectrum, including sites like cnn.com and nytimes.com.

3.3 Approximating the Partisanship of Subreddits and Users

To approximate the political partisanship of subreddits and Reddit users, we adopt the neural embedding approach based on the Word2Vec algorithm described by Waller et al. [159, 160], which learns subreddit and user embeddings/vectors based on the interaction data of users within subreddits. We illustrate this methodology in Figure 2. This is such that a high cosine similarity between two users would indicate that the two users comment/post in similar or the same subreddits; conversely, a high similarity between two subreddits would indicate that they share similar user bases. By computing subreddit and user similarity scores along a political partisanship dimension created when training the Word2Vec model, as in Waller et al. [160], this approach enables the approximation of the partisanship of users and subreddits. We utilize this approach as it allows us to avoid biases in previous manual labels of the political orientation of subreddits. Specifically, as in Waller et al. [160], we apply the Word2Vec algorithm to our Reddit data where subreddits are treated as "words" and users are treated as "contexts". In this approach, every individual instance of a Reddit user commenting or submitting in a given subreddit is considered a word-context pair. Upon aggregating these word-context pairs, we subsequently train a Word2Vec using skip-gram with negative sampling, outputting the vector embedding for each subreddit and for each user.

From our vector embeddings, as specified by Waller et al. [160], we identify the political partisanship dimension to then categorize the political orientation of individual subreddits and users. More concretely, after extracting our subreddit and user embeddings, we identify two similar communities that differ primarily in the dimension of interest; in this case, where we wish to measure partisanship, r/democrats and r/conservative. From the Word2Vec embeddings $\mathbf{sr}_{r/democrats}$ and $\mathbf{sr}_{r/conservative}$ that we elicited from these subreddits, we then compute the political partisanship dimensional vector $\mathbf{pr}_1 = \mathbf{sr}_{r/democrats} - \mathbf{sr}_{r/conservative}$. To ensure that the political dimension that we are studying is not overly specific to our seed communities of $\mathbf{sr}_{r/democrats}$ and $\mathbf{sr}_{r/conservative}$, we subsequently identify other pairs of similar communities whose difference vector has a high cosine similarity to our political partisanship dimensional vector $\mathbf{pr}_1$ (*i.e.*, other pairs of communities that differ primarily in political partisanship dimension; see Figure 2). For example, in our

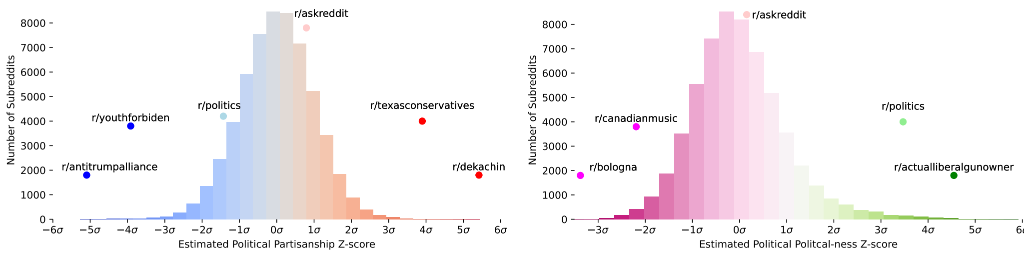


Fig. 3. Subreddit political partisanship and politicalness distribution — We determine the political partisanship (where a subreddit falls on the US left/right political spectrum) and how political a subreddit is by utilizing Waller et al.'s [160] method for creating subreddit and user embeddings via an extension of Word2Vec [97].

work, other pairs of communities that differed primarily along our political dimension included: r/liberalgunowners and r/gunpolitics, r/climatechange and r/climateskeptics, and r/askaliberal and r/askaconservative. As in Waller et al. [160], we average the vectors to get our final partisanship dimensional vector $\mathbf{pr}_1 = \mathbf{sr}_{r/democrats} - \mathbf{sr}_{r/conservative}$ using 10 unique political pairs:

$$\mathbf{pr} = \frac{1}{10}\sum_{i}^{10}\mathbf{pr}_i \tag{1}$$

To project individual subreddits and users onto the political partisanship dimension, we compute the cosine similarity between a given subreddit's or user's Word2Vec embedding $\mathbf{sr}_{r/any\_subreddit\_user}$ and the computed political partisanship dimension $\mathbf{pr}$ vector. To make these values more interpretable, as in Waller et al. [160], we determine the z-scores for the projected value on the political partisanship dimension. This is such that a community with a z-score of -1 could be interpreted as having a leftward stance with a political partisanship level of 1 standard deviation below the mean. As in Waller et al. [160], in addition to calculating the political partisanship of individual subreddits and users, by taking the sum of the vectors of our communities utilized to compute the political dimension, rather than the difference, we can also determine the "political"-ness of subreddits. By taking the sum, the "political"-ness score essentially measures the "strength of association" of each subreddit with the political partisanship dimension, assessing the level of political engagement of a community or user, rather than pinpointing their position on the political spectrum [160]. For example, the r/law subreddit, while not particularly partisan (-0.19σ), is over two standard deviations above the mean for politicalness (2.10σ). We showcase the resulting full distributions in Figure 3. We note that while we utilize this approach to approximate the partisanship and politicalness of different subreddits and users, this approach can be extended to other social dimensions. For example, within Waller et al.' [160] work, they identify an age axis by using the seed set of r/teenagers and r/RedditForGrownups.

We lastly note that given the many individual hyperparameters utilized within Word2Vec models (*e.g.*, embedding size, down-sampling threshold, starting learning rate, *etc...*), we perform a grid-search on these parameters and subsequently validate our outputted scores against the political partisanship scores of Waller et al. [160]. We select the model with partisanship scores that have the greatest Pearson correlation with those provided by Waller et al.[8] The partisanship scores utilized by Waller were validated against human labels as well as US census data, where they found that their calculated partisanship scores for subreddit-matched US communities had a moderate correlation

[8]We do not utilize the political partisanship scores provided by Waller et al. [160] given that their study is limited to 10,006 subreddits and given that they do not provide vectors or partisanship scores for individual users.

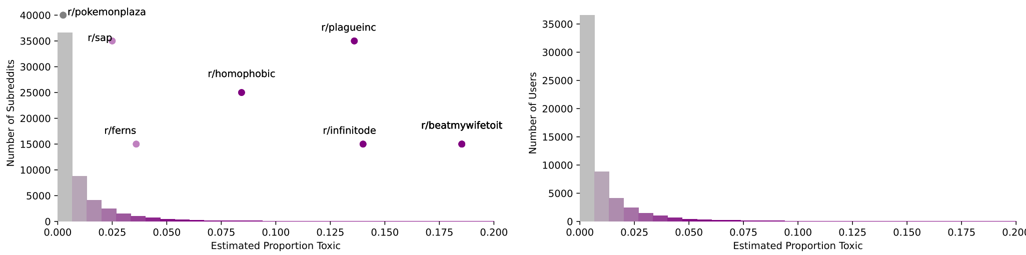


Fig. 4. Subreddit and User Toxicity scores—We determine the toxicity norms for subreddits with at least 100 comments and for users with at least 5 comments. Each user and subreddit has distinctive toxicity norms, posting toxic comments at different rates. At a threshold of 0.80, most comments/posts are not considered toxic by the Perspective API SEVERE_TOXICITY classifier.

with the Republican-Democrat vote differential in the 2016 presidential election (see Waller et al. [160] for additional details). Our final partisanship scores had a 0.828 Pearson correlation with the partisanship scores provided by Waller. We detail the hyperparameters and the values that we optimize over in Appendix A.

3.4 Identifying Toxic Comments and Approximating User and Subreddit Toxicity

To approximate the toxicity of Reddit users and subreddits, we utilize the Perspective API, a set of out-of-box toxicity classifiers from Google Jigsaw [86] that has been utilized extensively in prior works [96, 125, 136]. Each classifier takes comments as input and returns a toxicity score of 0.00–1.00; the closer a comment's score is to 1, the more likely the comment is to be toxic. In line with prior work, to consider a comment as toxic, we utilize a threshold of 0.80 on the `SEVERE_TOXICITY` classifier [29, 99]. As found by Kumar et al. [95, 96], utilizing this particular classifier, while limiting recall, provides an acceptable precision for identifying toxic online content. Finally, to ensure that the user and subreddit characteristics that we extract are robust, we only calculate average toxicity statistics for subreddits with at least 100 comments and users who posted at least 5 comments. We showcase the toxicity distributions for subreddits and users in Figure 4.

3.5 Ethical Considerations

Within this work, we focus on identifying trends in how 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 analyze specific users, we do not publish their usernames, and we do not attempt to contact or deanonymize any of them. We note that the Reddit submissions and comments analyzed in this work were public and available through the Pushshift API [14] at the time of collection [35].

4 Toxicity and Partisanship in Misinformation Posts

In this section, we examine the relationship between Reddit submissions utilizing unreliable information sources and their corresponding partisanship, toxicity, and user engagement (*i.e.*, number of comments). Using reliable news submissions as a control, we measure whether Reddit posts that link to known unreliable information sources predict increased toxicity. Within this section, we find that links to unreliable websites are correlated with increased toxicity and are posted more often on the right-leaning subreddits. After examining the distributional differences in several characteristics amongst the users and subreddits of unreliable and reliable news submissions,

Top Unreliable	# Links	Top Reliable	# Links	Top Unreliable Subreddits	# Links	Partisanship	Top Reliable Subreddits	# Links	Partisanship
oann.com	188,678	nytimes.com	493,032	r/TheNewsFeed	133,600	0.87σ	r/AutoNewspaper	1,010,948	-0.91σ
dailymail.co.uk	110,491	cnn.com	392,392	r/ConservativeNewsWeb	64,565	0.95σ	r/politics	426,931	-0.92σ
rt.com	27,347	reuters.com	245,633	r/OneAmericaNews	54,138	0.96σ	r/news	208,612	1.6σ
wnd.com	25,732	thehill.com	219,826	r/trendandstyle	47,171	-0.20σ	r/worldnews	195,644	1.38σ
newsmax.com	25,204	cnbc.com	179,157	r/StateoftheUnionNONF	27,232	-0.94σ	r/Coronavirus	178,555	0.20σ
americanthinker.com	22,247	nbcnews.com	174,430	r/Conservative	22,859	6.1σ	r/nofeenews	92,815	-0.86σ
sputniknews.com	19,736	yahoo.com	164,489	r/StonkFeed	16,941	-0.25σ	r/nytimes	89,795	-1.7σ
rumble.com	17,172	usatoday.com	147,323	r/TheBlogFeed	15,543	-0.65σ	r/NoFilterNews	85,960	-0.56σ
zerohedge.com	15,409	washingtonpost.com	128,579	r/conspiracy	13,510	3.2σ	r/NBCauto	83,361	-1.8σ
bitchute.com	12,788	latimes.com	124,742	r/boogalorian	8,730	2.3σ	r/CNNauto	79,436	-2.1σ

Table 1. Top reliable and unreliable websites hyperlinked within Reddit submissions and the top subreddits with unreliable websites and reliable websites hyperlinked. Altogether, within our set of studied 57K subreddits, we identify 633,585 submission hyperlinks to our set of unreliable news websites and a total of 7,546,917 submission hyperlinks to our set of reliable news websites.

we finish this section by fitting a linear model and a negative binomial model to understand the degree to which each of these features predicts toxicity and user engagement on Reddit.

Setup. We identify submissions that link to our 1,054 unreliable and 3,754 reliable websites. Altogether, we find 633.59K submissions of unreliable news websites and a corresponding set of 5.29 million comments; we further find 7.55 million submissions that link to our set of reliable websites and 267 million corresponding comments. We list the most frequently linked websites and subreddits that most commonly link to our sets of sites in Table 1. Altogether, hyperlinks to unreliable websites were posted in 9,462 subreddits and to reliable websites in 29,673 subreddits (8,611 subreddits had links to both). The difference in the magnitude of submissions is likely due to the greater popularity and widespread appeal of reliable mainstream news compared with alternative, fringe websites [67], in addition to our larger list of reliable news websites. Indeed, utilizing the Alexa Top Million list from March 1, 2021 [5], we find that 991 reliable news websites (26.39%) were in the top 100K websites compared to 139 unreliable websites (13.19%).

For the rest of this section, while using partisanship, politicalness, and toxicity averages computed from our full Reddit dataset (see Section 3), we analyze the set of Reddit submissions and Reddit comments that involve unreliable and reliable website submissions. We note that in our analysis, we remove "AutoModerator" comments and comments from accounts labeled as "bots" (*e.g.*, u/RemindMeBot).

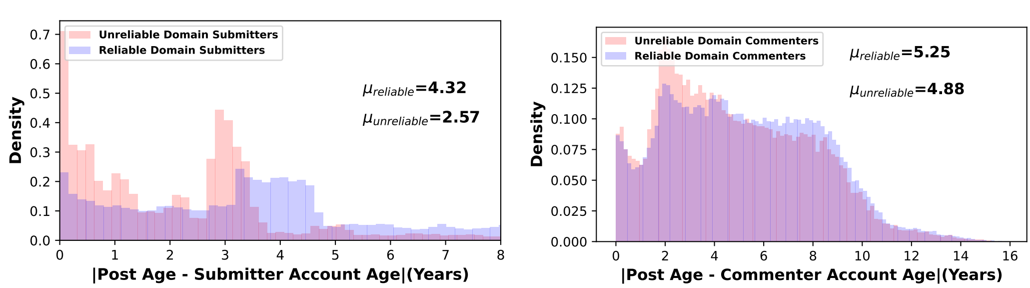
4.1 Differences Between Unreliable and Reliable Website Submissions

Across our dataset, we find that 1.26% of all comments within our datasets were classified as toxic (*i.e.*, Perspective `SEVERE_TOXICITY` score >0.80), 1.24% of comments under reliable website submissions were considered toxic, and 1.55% of comments on unreliable submissions (a 25% relative increase). However, as previously mentioned, these comments are largely posted in different communities on Reddit and likely by different users. Performing a comparison across individual subreddits, we find that there remains a mean absolute percentage increase of 0.35% (32.2% relative increase) in toxicity ($p < 1 \times 10^{-16}$) for toxicity on unreliable news articles compared to reliable news articles. In this section, we thus determine the differences between subreddits and users who interact with reliable versus unreliable news to understand this increase in toxicity.

Subreddits. As seen in Table 2, on average, the subreddits where unreliable website submissions are posted are 1.13 standard deviations more right-leaning on the US political spectrum than those of reliable websites. This accords with previous research that has found that right-leaning users and ecosystems are more likely to spread misinformation [88]. However, we also observe that unreliable website submissions tend to be posted in subreddits that are typically 0.75 standard

	Unreliable	Reliable	Cohen's D
Avg. Subreddit Partisanship	0.96σ	-0.17σ	0.79
Avg. Subreddit Politicalness	2.37σ	3.12σ	-0.47
Avg. Subreddit Toxicity	2.01%	1.40%	—
Avg. Submitter Partisanship	-0.04σ	-0.19σ	0.19
Avg. Submitter Politicalness	-0.01σ	0.49σ	-1.42
Avg. Submitter Toxicity	0.93%	0.90%	—
Avg. Submitter Account Age (Years)	2.57	4.32	-0.54
Avg. Commenter Partisanship	0.56σ	0.09σ	0.57
Avg. Commenter Partisanship Var.	0.45	0.48	-0.07
Avg. Commenter Politicalness	0.20σ	0.26σ	-0.20
Avg. Commenter Politicalness Var.	0.13	0.15	-0.19
Avg. Commenter Toxicity	1.48%	1.36%	—
Avg. Commenter Account Age (Years)	4.88	5.25	-0.12
% Removed Comments	2.01%	2.82%	—
% Mod/Admin Involved	16.74%	16.26%	—

Table 2. We determine different characteristics of the subreddits, submitters, and commenters that interact with reliable and unreliable website submissions and subsequently determine the Cohen's effect sizes between unreliable news submissions and reliable news submissions for these characteristics. We perform Mann-Whitney U tests to ensure that the differences in the averages between unreliable and reliable website submissions are significant. We perform two-sample proportion tests for the percentages. Note, even though after we applied a Bonferroni correction, all p-values tested were $p < 1 \times 10^{-16}$ and significant. We identified marked differences between unreliable and reliable Reddit submission threads, with unreliable submission threads being more right-leaning, having higher toxicity, being posted by more right-leaning users, being posted by younger accounts, and having more toxic comments.



(a) Account Age of Submitters (b) Account Age of Commenters

Fig. 5. Younger accounts are much more likely to submit and comment on unreliable website submissions.

deviations less political than reliable website submissions. For example, r/StreetFighter, a subreddit dedicated to the video game Street Fighter (politicalness=-0.47σ), contained 409 submissions to 4chan.org and r/MMA (politicalness=-0.43σ) had 81 hyperlinks to the known Russian propaganda website rt.com [23] and the far-right conspiracy website infowars.com [155]. We further find that unreliable website submissions tend to be in subreddits with higher average toxicity (2.01% v.s. 1.40% of comments), which may also explain the higher likelihood of toxic comments in response to misinformation posts.

Submitters. In line with prior work, we find that users who submit unreliable websites articles as Reddit submissions tend to be more right-leaning (-0.04σ v.s. -0.19σ), tend to be less political (-0.01σ v.s. 0.49σ), tend to have slightly more toxic comments (0.93% v.s. 0.90%), and tend to have younger accounts (Table 2). We note, as mentioned before in our calculation of user toxicity, that we utilize the submitters' Reddit comments to determine their average toxicity. Performing a subreddit pairwise comparison (*i.e.*, comparing the users who submitted unreliable websites in one subreddit to the users who also submitted reliable websites in the *same* subreddit), we indeed find that users who submit unreliable websites tended to be more right-leaning (Cohen's D = 0.26, $p < 1 \times 10^{-16}$ using the paired Wilcoxon signed-rank test), were very slightly more political (Cohen's D = 0.01, $p < 1 \times 10^{-16}$), and were slightly more toxic overall (0.12% absolute percentage increase, $p < 1 \times 10^{-16}$). We further observe that submitters of unreliable website hyperlinks tend to have younger accounts. As argued elsewhere, when posting inflammatory, revealing, or otherwise sensitive information [9, 95, 100], Reddit users often utilize disposable "throw-away" accounts that are used only to post this information anonymously. Indeed, as seen in Figure 5, within our dataset, we find that while only 0.88% of reliable website submissions are submitted within the first week of an account's lifespan, 2.64% are submitted in the first week for unreliable websites (we perform a proportion test and find this difference to be significant $p < 1 \times 10^{-16}$).

Commenters. Commenters on unreliable website submissions tend to be slightly more right-leaning (0.56σ v.s. 0.09σ), slightly less political (0.20σ v.s. 0.25σ), and slightly more toxic (1.48% v.s. 1.36%). Performing a subreddit pairwise comparison (*i.e.*, comparing the users who commented on unreliable websites in one subreddit to the users who commented on reliable websites in the *same* subreddit), we find, however, that the users who comment on unreliable websites have no significant difference (using the paired Wilcoxon signed-rank test) in partisanship nor toxicity, but do differ slightly in politicalness (Cohen's D = -0.07). We thus see that after accounting for the subreddit, *that is largely the same type of users who comment on unreliable and reliable website submissions within a given subreddit*. Despite seeing that within subreddits the users of similar partisanship and toxicity comment both on unreliable and reliable news submissions, again performing this subreddit pairwise comparison, we find as previously reported that there is a mean absolute percentage increase of 0.35% (32.2% relative increase) in toxicity ($p < 1 \times 10^{-16}$) for unreliable submissions within each subreddit. This illustrates that, despite similar users participating in conversations surrounding unreliable and reliable news within a given subreddit, comments on unreliable news tend to have more toxic language. As for submitters (Figure 5), we find that commenters on unreliable website submissions have younger accounts than those for reliable website submissions (4.88 years v.s. 5.25 years). We note that, as with submitters, this may partially explain the increased toxicity of unreliable submission commenters.

Moderation and Removed Comments. A potential confounder that can cloud our analysis is the activity of Reddit moderators. Reddit moderators are members of particular subreddit communities who help set rules and norms as well as help moderate content [8]. When a moderator on the Reddit platform removes a comment, the comment is replaced with "[removed]" and other Reddit users can no longer view the comment. As seen in Table 2, on average, reliable website submissions are more moderated compared to unreliable website submissions (with an average of 2.01% comments for unreliable submissions being removed compared to 2.82% for reliable submissions). However, again performing a subreddit-wise pairwise comparison, we find that within the subreddits where both reliable and unreliable submissions appear, unreliable news commenters are moderated more heavily (Cohen's D = 0.37, $p < 1 \times 10^{-16}$ using the paired Wilcoxon signed-rank test). This indicates, within subreddits that have both unreliable and reliable domain hyperlinks, that unreliable ones are moderated more heavily; conversely, outside of these subreddits, these unreliable website submissions are moderated more leniently. For example, within the r/bicycling subreddit, while 2.01% of reliable domain comments were removed, 19.35% of unreliable domain comments were removed. In contrast, within the r/bitchute, where there were no comments on reliable news domain hyperlinks, only 0.50% of comments were removed (BitChute is an alternative to YouTube known for hosting toxic and conspiratorial content [151]).

Variable	**Coefficient**	**Std.**
Commenter Toxicities	$4.99 \times 10^{-2***}$	1.18×10^{-2}
Subreddit Partisanship	$6.64 \times 10^{-3**}$	2.24×10^{-3}
Subreddit Toxicity	$2.33 \times 10^{-2**}$	9.03×10^{-3}
Removed Comments	$1.81 \times 10^{-2*}$	8.07×10^{-3}
Moderator or Admin Involved	$1.64 \times 10^{-2*}$	8.15×10^{-3}

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

Table 3. Model of the toxicity of the comments in Reddit submissions. We fit a linear model with LASSO regularization to model the percentage of toxicity in each of the Reddit threads that contained a reliable domain or an unreliable domain in the submission. As seen above, within Reddit threads, we find that as subreddits become more right-leaning and have higher average toxicity levels, this predicts more toxicity within individual Reddit threads.

We lastly examine the cases where a moderator left a comment or interacted with users in the subreddit threads. We utilize the "distinguished" moderator/admin flag supplied in our Pushshift dataset to identify moderators. As seen in Table 2, across all our submissions, moderators were involved in slightly more submissions in unreliable domain submissions, either as the submitter or as a commenter. We find that the comments of submissions that had a moderator/admin involved were less toxic than those that did not (1.04% v.s. 1.66% for unreliable website submissions; 0.70% vs 1.12% for reliable website submissions). Performing a subreddit-wise pairwise comparison on the proportions of submissions per subreddit that had moderator involvement, we again see that unreliable news websites were very slightly more likely to have a moderator involved (Cohen's D =0.05, $p < 1 \times 10^{-16}$ using the paired Wilcoxon signed-rank test).

Summary. In this section, we showed that links to websites known to spread unreliable information are correlated with more toxic comments: toxic comments under unreliable website submissions are posted at a rate of 1.55% while toxic comments in response to reliable website submissions are posted at a rate of 1.24%. Within individual subreddits, we find that on average, the toxicity rate increases on average by an absolute 0.35% (32.2% relative increase). In addition, we observed that users who post and comment on misinformation are right-leaning.

4.2 Prediction of Toxicity By Commenter and Subreddit Characteristics

Having seen the higher toxicity in response to various subreddit thread conditions, we now examine how the factors previously examined interact with one another to collectively predict toxicity.

Setup. We fit a linear model with LASSO regularization to identify the most predictive factors of toxicity in Reddit comment threads. Our dependent variable is the percentage of toxic comments in each thread, derived from 633.59K submissions linking to unreliable domains and 7.55M submissions linking to reliable domains. LASSO enables variable selection by shrinking less informative coefficients to zero, helping prevent overfitting and focusing the model on the most meaningful predictors [126]. Table 3 lists the six variables retained in the final model, along with their coefficients and standard errors.

Results. As shown in Table 3, our LASSO model identifies five variables as significant predictors of toxicity in Reddit comment threads. The strongest predictor is the average toxicity of the commenters who respond to a submission ($\beta = 4.99 \times 10^{-2}$), consistent with our earlier findings that users who have exhibited toxic behavior tend to perpetuate toxic discourse in other threads. We further observe that subreddit characteristics also play a substantial role. The model finds that higher overall subreddit toxicity is associated with an increase in thread-level toxicity ($\beta = 2.33 \times 10^{-2}$), reaffirming that toxic community norms influence individual behavior. Additionally, we find that increased subreddit partisanship is also associated with elevated toxicity ($\beta = 6.64 \times 10^{-3}$). This aligns with our findings in the previous section, which found that the unreliable website Reddit submissions, which have higher levels of toxicity, are also more right-leaning. Moderation signals again appear as predictive factors for toxicity, as found in the previous section. However, unlike our previous finding, we observe that the proportion of removed comments in a thread is positively associated with toxicity ($\beta = 1.81 \times 10^{-2}$), as is the involvement of a moderator or admin ($\beta = 1.64 \times 10^{-2}$). These results are also consistent with the interpretation (after taking into account other factors) that moderation activity tends to follow or respond to toxicity rather than prevent it [150].

Notably, variables such as commenter partisanship, commenter politicalness, subreddit politicalness, and the reliability of the linked domain were not selected by the LASSO model. This suggests that these factors did not provide additional explanatory power beyond the behavioral and subreddit-level features retained in the final model. Having observed this phenomenon at a Reddit thread level, in further sections, we examine if there are any substantial differences in how users interact with one another and on Reddit when users comment on unreliable website submissions.

Summary. After fitting a LASSO-regularized regression model, we find that the toxicity of individual commenters, the partisan alignment and toxicity of subreddits, and moderation-related indicators are key predictors of toxic discourse in Reddit threads. In contrast, factors such as news source reliability and user political orientation were excluded, indicating that toxicity within individual norms may be more strongly driven by community norms and individual behavioral norms.

4.3 Prediction of Engagement via Toxicity, Use of Unreliable Sources, and Partisanship

Having shown that unreliable sourced submissions have increased toxicity on Reddit, but that this increased toxicity is largely accounted for by other subreddit and user characteristics, we now determine what factors induce engagement with unreliable information. Namely, are toxicity, politicalness, and partisanship correlated with more engagement with misinformation?

Setup. To measure user engagement with unreliable and reliable website submissions, we utilize the number of comments that each submission receives.[9] As before, to properly model the number of comments, we remove comments from Reddit's "AutoModerator" accounts or explicitly "bot" labeled accounts. Altogether, we analyze our set of 633.59K unreliable website submissions, our set of 7.55M reliable website submissions, and each of these sets' associated comments.

To model the number of comments on submissions, we utilize a zero-inflated negative binomial regression [129]. Within our model, each observation data point represents a single submission and its associated number of posted comments. We utilize a zero-inflated negative binomial regression as it appropriately models our set of count data — some submissions garner thousands of comments while others garner only a few. 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 [110]. We further utilize the zero-inflated version of this regression, given the heavy preponderance of submissions that do not receive any comments.

We finally note that zero-inflated negative binomial regressions return two sets of coefficients. One set of coefficients, the zero-inflated coefficients, estimated using logistic regression, reports the probability that the given submission would receive zero comments as a function of the covariates. Positive coefficients for these zero-inflated coefficients indicate that increases in the predictor variable make the submissions receiving zero comments more likely. Thus, the more negative a

[9]We utilize the number of comments rather than the number of upvotes/downvotes because Pushshift often fails to keep up-to-date information about the number of votes for submissions [14].

coefficient, the more the given covariate correlates with inducing at least 1 comment. The second set of coefficients, the negative binomial coefficients, models the number of comments as a function of the covariates. For these coefficients, the more positive the coefficients, the more comments the submission was likely to have received as a function of the covariate. 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. As factors influencing the number of comments, we utilize:

(1) the submitter's admin/moderator status
(2) the relative age of the account that posted the submission
(3) the submitter's partisanship
(4) the submitter's politicalness
(5) the submitter's account's age
(6) the submitter's toxicity
(7) the subreddit's partisanship
(8) subreddit's politicalness
(9) the subreddit's toxicity
(10) the average number of comments per submission of the subreddit

We employ backward variable selection based on AIC to select variables. We utilize backward variable selection as it allows us to consider the full model with all variables while also allowing us to create a final parsimonious model [108]. We note the limitation, however, that the performance and stability of our backward elimination approach can be sensitive to the presence of weak predictors or collinearity, although this is less of a risk in cases like ours, with a large number of datapoints relative to the number of regression coefficients [154].[10]

Results. We now give an overview and describe some of the implications of our results using our negative binomial regression to predict levels of user engagement based on levels of politicalness, partisanship, toxicity, and the use of unreliable news articles.

Submitter Admin/Moderator Status. For unreliable website submissions, we find that when a moderator posts the submission, they are more likely to get at least one comment compared to a non-moderator account ($\beta = -1.25$). In contrast, for reliable website submissions, we find that these moderator or admin accounts are less likely to gain at least one comment compared to non-moderator accounts ($\beta = 0.20$). For both unreliable and reliable website submissions, however, we observe that when admin or moderator accounts' posts do gain comments, they are more likely to receive more comments than normal accounts. This largely accords with moderators' role on the platform in making announcements within subreddits [103].

Submitter Toxicity. We note, as mentioned before in our calculation of user toxicity, that we utilize the submitters' Reddit comments to determine their average toxicity. Examining the submitting users' toxicity, we see somewhat similar behaviors for both reliable and unreliable information submissions. Most notably, as the submitting users become more toxic, for both unreliable and reliable website submissions, they are more likely to receive at least one comment. However, for unreliable website submissions, we observe that the submitter's toxicity has a much larger effect

[10]We spot-check our results to ensure that the higher the average number of comments in a given subreddit, the more likely a submission is to see comments *and* that this average correlates with more comments on submissions. In other words, we check that submissions in subreddits where users comment more are also predicted to receive more comments. As seen in both Tables 4 and 5, for both unreliable and reliable website Reddit submissions, as the average number of comments in a subreddit increases, (1) the more likely a submission is to receive comments and (2) the more comments it is likely to receive.

on the probability of receiving at least one comment ($\beta = -19.47$ v.s. $\beta = -0.06$). This illustrates that while for unreliable websites, increased toxicity may receive greater initial engagement, this effect is not as strong for reliable websites. However, again in both cases, we see that while user toxicity often leads to one other user commenting, we see that this toxicity often does not lead to more comments (the coefficient for unreliable websites is not statistically significant).

Number of Comments on Unreliable Website Submissions				
	Zero-Inflated negative coefficient = more likely to get comments	Std. Error	Negative Binomial positive coefficient = more comments	Std. Error
Intercept	3.30^{***}	0.14	0.36^{***}	0.04
Submitter Is Moderator	-1.25^{***}	0.06	0.30^{***}	0.04
Submitter Toxicity	-19.47^{***}	1.73	-0.15	0.54
Submitter Politicalness	-1.14^{***}	0.24	-0.49^{***}	0.09
Submitter Partisanship	5.41^{***}	0.34	-0.12	0.12
Submitter Age	-0.23^{***}	0.01	0.02^{***}	0.003
Subreddit Toxicity	-0.99^{***}	0.03	-0.09^{***}	0.01
Subreddit Politicalness	1.05^{***}	0.03	0.04^{***}	0.01
Subreddit Partisanship	0.64^{***}	0.03	0.12^{***}	0.01
|Subreddit Partisanship - Submitter Partisanship|	-0.19^{***}	0.04	-0.34^{***}	0.01
Average # Subreddit Comments	-2.48^{***}	0.05	0.12^{***}	0.001

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

Table 4. Fit of our zero-inflated negative binomial regression on the number of comments on our set of unreliable URL submissions across different subreddits. As seen above, we find that moderator submissions are more likely to receive comments. We further observe that as Reddit submitters' toxicity and politicalness increase, their submissions are more likely to receive initial comments. We finally observe that as a subreddit's average toxicity increases, unreliable website submissions are more likely to receive initial comments.

Number of Comments on Reliable Website Submissions				
	Zero-Inflated negative coefficient = more likely to get comments	Std. Error	Negative Binomial positive coefficient = more comments	Std. Error
Intercept	-3.37^{***}	0.02	0.63^{***}	0.01
Submitter Is Moderator	0.20^{***}	0.01	0.49^{***}	0.01
Submitter Toxicity	-0.06^{***}	0.003	-0.02^{***}	0.002
Submitter Politicalness	2.43^{***}	0.02	0.22^{***}	0.006
Submitter Partisanship	-0.31^{***}	0.006	-0.05^{***}	0.003
Submitter Age	-0.15^{***}	0.004	0.12^{***}	0.002
Subreddit Toxicity	0.23^{***}	0.005	0.11^{***}	0.004
Subreddit Politicalness	0.79^{***}	0.004	0.19^{***}	0.002
Subreddit Partisanship	0.51^{***}	0.004	0.43^{***}	0.002
|Subreddit Partisanship - Submitter Partisanship|	0.46^{***}	0.006	0.08^{***}	0.003
Average # Subreddit Comments	-2.94^{***}	0.01	1.60^{***}	0.003

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

Table 5. Fit of our zero-inflated negative binomial regression on the number of comments on our set of mainstream URL submissions across different subreddits. As seen above, we find that moderator submissions are less likely to receive initial comments. We further observe that as the Reddit submitter's toxicity and partisanship increase, their submission is more likely to receive initial comments. However, in contrast to unreliable website submissions, highly political submitters are less likely to receive initial comments on their submissions. We finally observe that as a subreddit's average toxicity increases, while reliable website submissions are less likely to receive comments, they are more likely to receive increased amounts of comments when they receive at least one.

Submitter's Politicalness. While we observe that for unreliable websites, the higher a user's politicalness, the more likely to receive at least one comment ($\beta = -1.14$), there is the opposite effect for reliable websites ($\beta = 2.43$), where they are less likely to get comments the higher the submitter's politicalness. This appears to indicate that in the case of reliable website submissions, other Reddit users are perhaps being "turned off" and are engaging less with highly ideological users compared to less political users [73]. However, we also find that the more political a user becomes (if the submission gets comments), the fewer comments unreliable website submissions are likely to receive ($\beta = -0.49$), in contrast to reliable website submissions, which receive more comments ($\beta = 0.22$). This illustrates that highly politicized users may be more likely to engender a discussion amongst users for reliable website submissions, but are less effective at gathering comments for unreliable website submissions.

Submitter Partisanship. For unreliable websites, we find that the more right-leaning a user is, the less likely the user's post is to attract any user comments. Given the right-leaning nature of most of the subreddits (0.97σ) in which unreliable domain posts are submitted, this could likely be due to these users' posts being seen as "normal", being unnoticed, and the posts not receiving many comments ($\beta = 5.41$). In contrast, for reliable news ($\beta = -0.31$), we see that as the submission's submitter becomes more politically right-wing, the more likely their posts are to receive comments. Given reliable website submissions tend to be posted in left-leaning subreddits (-0.17σ), submissions from more right-leaning users may be seen as more novel, resulting in at least one user comment [76, 91]. This also supports prior research that has found that out-group animosity may drive online engagement [127]. However, despite right-leaning users being able to attract at least one comment for reliable website submission, we also observe that as the posting user becomes more right-leaning partisan ideological, the fewer comments their post is likely to receive overall ($\beta = -0.05$) [73].

Submitter Age. For both unreliable and reliable websites, we find that older accounts are more likely to receive at least one comment, and that the older the account, the more comments that its submission is likely to get. This may indicate that accounts with more history may attract more engagement with their posts.

Subreddit Toxicity. Looking at the subreddit toxicity coefficient in predicting whether a submission receives comments, we see a marked difference between reliable website submissions and unreliable website submissions. We see, notably, for misinformation submissions, the more toxic a subreddit is, the more likely the submission is to get comments ($\beta = -0.99$). In contrast, for reliable website submissions, the more toxic the subreddit, the more likely the submission is to not get any comments at all ($\beta = 0.23$). 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 battleplan.news [145] published a report entitled *"Wake Up! Even The Masks Made You Sick"!*[11] and these types of posts in more toxic environments appear to garner more initial engagement.

However, we further find, for reliable website submissions, that as subreddit toxicity increases, the more comments submissions are likely to garner ($\beta = 0.11$). In contrast, for unreliable website submissions, the more toxic the subreddit, the fewer comments the submission is likely to garner ($\beta = -0.09$). This reflects that *when* reliable website submissions get noticed or spark engagement in a toxic community, the more toxic the environment, the more users seem to comment and engage with the submissions. In contrast, when articles from unreliable sources are noticed in toxic environments, they do not appear to draw extensive interactions. We thus see that reliable website

[11] https://web.archive.org/web/20220801105629/https://battleplan.news/watch?id=62cf06f3c0f117796a9553b7

submissions are more often ignored in toxic subreddits compared to unreliable website submissions; simultaneously, we observe that as communities get more toxic, they tend to comment more on reliable information and less on unreliable information submissions.

Subreddit Politicalness. For both unreliable and reliable website submissions, the more political a subreddit, the fewer users are likely to comment at all (β = 1.05 and β = 0.79. This largely may demonstrate a novelty aspect, given that highly political subreddits receive constant news updates. However, for both unreliable (β = 0.04) and reliable submissions (β = 0.19), we find that when a submission is commented on, subreddits' politicalness increases the likelihood of more comments. This association is again probably largely a result of the fact that subreddits that are more politically engaged on the US political spectrum are more likely to be interested in news [59] and subsequently comment more on posts when they gain traction.

Subreddit Partisanship. We find that for reliable websites, the more politically right-leaning a subreddit, the less likely it is to gain any comments (β = 0.51). Rather, as documented by Wang *et al.* [161], subreddits like these often ignore more trustworthy sources. We similarly find that for unreliable websites, the more politically right-leaning, the less likely these posts are to get any comments (β = 0.64). As before, given the right-leaning nature of most of the subreddits (0.97σ) in which unreliable domain posts are submitted, this could likely be due to these users' posts being seen as "normal." In contrast, for both misinformation and reliable website submissions, we find that as the subreddit's right-leaning partisanship goes up, the more comments given submissions are likely to garner.

|Subreddit Partisanship - Submitter Partisanship|. For unreliable websites, we find that as the difference between the submitter's partisanship and the subreddit's partisanship increases, the more likely the post is to get at least one comment (β = −0.19). Various works have found that users not aligned to political norms of a given environment [125] often provoke engagement from users as they become "outraged" by the presented content [52, 91], and this can largely be observed here. We note that we do not observe a similar phenomenon for reliable website submissions (β = 0.43), which may result from the reliable website submission being seen as "bland" regardless of who posts them and being unable to receive initial comments. However, for reliable website submissions, we find that as the difference between the submitting user's partisanship and the subreddit's partisanship increases, the more comments the submission is likely to get. This indicates that when the reliable submission manages to gain initial comments, the further the submitter's partisanship for the subreddit as a whole, the longer the ensuing conversation. In contrast, for unreliable website submissions, our model finds that as a submitter's partisanship moves further away from the subreddit's own partisanship, after initially accruing an initial comment, it is less likely to gain additional comments.

4.4 Summary

In this section, we find that submitter toxicity, submitter politicalness, submitter age, and subreddit toxicity all encourage initial interaction with unreliable website submissions. In contrast, submitter toxicity and subreddit toxicity play much more muted roles for reliable news submissions, with the subreddit toxicity actually predicting less initial engagement with reliable news sources. This appears to overall suggest a higher degree of initial engagement with unreliable news outlets in political and toxic settings compared to reliable sources.

We further find that moderator involvement, subreddit politicalness, and subreddit partisanship all encourage users to have longer sustained interactions with unreliable information, while subreddit toxicity predicts shorter conversations. In contrast, for reliable news, we find that subreddit toxicity, subreddit politicalness, and subreddit partisanship all predict increased user engagement (if users initially comment at all). This illustrates that while toxic environments may receive initial engagement with unreliable news, it does not predict sustained interactions, with the opposite being true of reliable news.

5 Unreliable Websites and Polarized Toxic Interactions

In the previous section, we showed that unreliable website submissions are correlated with increased toxicity and that increased toxicity is also correlated with comments on unreliable website submissions. To understand the user-level dynamics of toxicity in response to unreliable news submissions, we reconstruct the conversational dyads that exist underneath each Reddit submission. We note that these dyads, as in Kumar et al. [95], reflect the individual interactions between two user accounts within the comments in Reddit threads. Using the approach outlined in Section 3.1, we then determine the partisanship, politicalness, and average toxicity of the users in these conversational dyads, mapping out different types of political interactions. From these averages, we label users as right-leaning (greater than 0.5σ partisanship) or left-leaning (less than -0.5σ partisanship). Then, looking at each conversational dyad, we determine if each comment is toxic using the Perspective API (as outlined in Section 3.1). As an example of such a dyad, in the r/Coronavirus subreddit, a user with a left-leaning bias posted:

> *Why oh why are people spitting on strangers? And can we get some spit for the evil 80 who own half the planet? No? Ok.*

to which another user with a right-leaning bias replied:

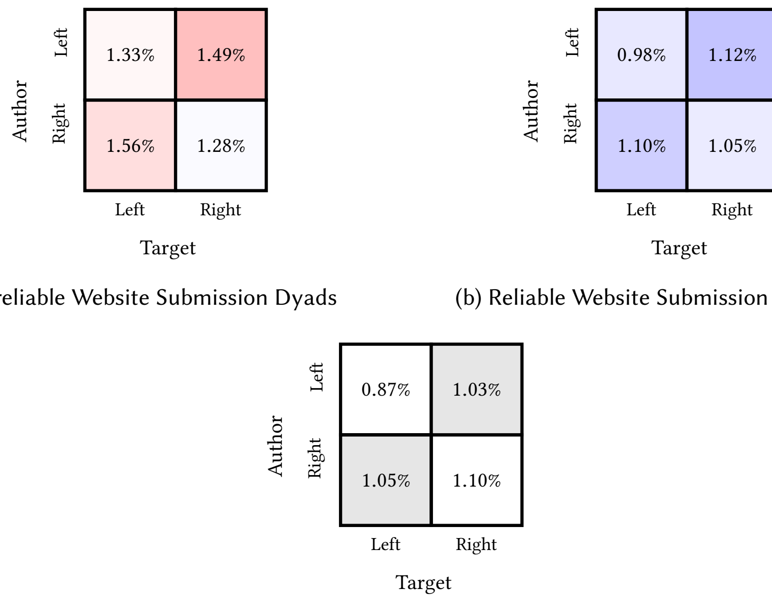
> *Come the fuck on. I don't care what your opinions are or if it was just a really shitty joke. Don't wish for people to catch this, that's an asshole move right there.*

For a comparison of how conversations differ between unreliable website and reliable website comments, we finally separate the set of conversational dyads that appear under unreliable versus reliable website submissions. From this comparison, in this section, we find that interactions between users of differing political orientations have increased toxicity, with this being particularly true in Reddit threads under unreliable news Reddit submissions. We note that within this section, for clarity, for our derived dyads, we refer to "target" as the original commenter and "author" as the commenter who is responding to the target.

5.1 Interactions within Unreliable and Reliable Information Ecosystems

We observe (as expected) that many users primarily interact with users of the same partisanship [146]: 71.80% of interactions were between users who share the same partisanship-lean. For unreliable news submissions, this rises to 83%, and for reliable website submissions, it drops to 66%. We similarly find that 72.08% of toxic interactions (where a user responded to another user with a toxic reply) were between users who shared the same partisanship leaning among all dyads, 80.63% for unreliable website submission dyads, and 63.34% for reliable website submission dyads. The differences in percentage of homophily in user interactions are likely because, as previously found, unreliable domains are largely posted in somewhat more insular subreddits (average partisanship = 0.97σ; Table 2) and in communities with higher degrees of toxicity (2.01%; Table 2).

Despite users largely interacting with users of the partisanship, we find some increased rates of affective polarization between users of different partisanship. As seen in Figure 6, we observe increased toxicity between users of different partisanship for our set of website submissions, with this difference most marked for unreliable website submissions. Indeed, calculating the odds ratios between the percentages of inter-partisanship toxicity against those of intra-partisanship toxicity, we get values of 0.99 across all dyads, 1.19 for unreliable domain dyads, and 1.08 for reliable domain



(a) Unreliable Website Submission Dyads

(b) Reliable Website Submission Dyads

(c) All Submission Dyads

Fig. 6. Percentage of interactions that are toxic in all, unreliable, and reliable website submissions for right and left-leaning authors against right and left-leaning targets. Here "target" refers to the original commenter and "author" to the commenter who is responding to the target. We find that for website submissions, and most notably for unreliable website submissions, users of differing partisanship biases (*i.e.*, left v.s. right) had higher odds of leaving toxic comments to each other.

dyads. We thus observe a slight increase in inter-partisanship toxicity between users who comment under submissions with attached domain hyperlinks. Further, calculating the odds ratio between the rates of toxicity between unreliable websites and reliable website conversational dyads, we get values of 1.38 for inter-partisanship toxicity and 1.26 for intra-partisanship toxicity. We thus observe that amongst our set of conversations, there is an even further heightened rate of affective polarization for unreliable news interactions compared to reliable news interactions.

5.2 Modeling Toxic Interactions Between Users

To concretely show that users of different political stripes are more likely to reply in a toxic manner to each other in conversations under unreliable domain submissions, we fit our dyad data of toxic interactions into an exponential random graph model. An Exponential Random Graph Model (ERGM) is a form of modeling that predicts connections (*e.g.*, toxic interactions) between nodes (users) in a given network [80]. ERGM assumes that connections are determined by a random variable p^* that is dependent on input variables. As in Chen *et al.* [27] and Peng *et al.* [118], we utilize this modeling as it does not assume that its data input is independent and can model network interactions [80, 156].

Setup. Utilizing our ERGM and the conversational dyads that we constructed between users from the last section, we predict the probability of toxic interactions between two users within misinformation submissions as a function of:

(1) the total number of connections between all users/intercept
(2) the users' percentage of toxic comments

Unreliable Domain Interactions	Coeff.	Std. Error
Intercept/Edges	8.65^{***}	0.05
User Partisanship Differences	-0.20^{***}	0.04
User Toxicity	5.88^{***}	0.46
Shared Subreddits Comments	0.004^{*}	0.001
Reciprocity	4.79^{***}	0.18

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

Reliable Domain Interactions	Coeff.	Std. Error
Intercept/Edges	8.73^{***}	0.05
User Partisanship Differences	-0.29^{***}	0.04
User Toxicity	6.48^{***}	0.74
Shared Subreddits Comments	0.001^{***}	0.0004
Reciprocity	3.97^{***}	0.27

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

Table 6. Toxic Unreliable and Reliable Website Submission Interactions — As confirmed in our ERGM, differences in the political orientation of users are predictive of increased incivility and toxicity, with users of differing political orientations more likely to engage in toxic interactions within misinformation submissions than on mainstream submissions. Similarly, the higher each user's toxicity norm, the more they are likely to target other users with toxic comments.

(3) the users' partisanship
(4) the difference in the author and target's political partisanship
(5) the users' politicalness
(6) the account age of the two users
(7) the reciprocity between the two users (*i.e.*, if both users had a toxic comment aimed at each other)
(8) the number of comments that the two users have in subreddits in which they both post comments
(9) toxic triangle relationships between triads of users

We include the number of comments that the users have made in shared subreddits to account for the fact that users with more overlap in user activity (*i.e.*, frequenting the same subreddits) are more likely to interact with one another and thus have toxic interactions. We display the goodness of fit for our EGRM statistics (edges) using the variables in our model in Appendix B — we find, using the `statnet` R package, that our fitted models adequately reproduce key network structures, with Monte Carlo p-values indicating no significant deviation from the observed network.

Results. We find that account age, partisanship, and the politicalness of a given user do not have significant effects on the likelihood of toxic interactions — all were removed from the fitted models. This indicates that just because a user is highly partisan or political, it does not necessarily mean that they are likely to engage in toxicity. For all domain interactions, as seen in Table 6, we find that (1) that the more toxic a user, the more likely they are to engage in toxic interactions, and (2) that users are more likely to respond in a toxic manner to users who engage with them in a toxic manner (reciprocity). Indeed, we find that in unreliable website submission comments compared to reliable website submission comments, users are more likely to reply in a toxic manner to another user if that user has already corresponded with them in a toxic manner ($\beta = 4.79$ v.s. $\beta = 3.97$). However, most importantly, we find that while most toxic interactions occur among users who are politically similar to each other, compared to reliable domain interactions, users discussing unreliable website submissions are *more* likely to send toxic comments to users of different political ideologies than users under mainstream submissions ($\beta = -0.20$ v.s. $\beta = -0.29$).

Summary. In this section, we showed that unreliable website submissions not only promote higher levels of toxicity in their comments but are also correlated with increased inter-partisanship toxicity compared to reliable website submissions. Fitting an ERGM to our toxic conversational dyads posted in response to misinformation stories, we show that political differences, along with reciprocity and each user's toxicity, drive more toxic interactions.

6 Limitations

In this work, we used a quantitative, large-scale approach to understand the role of misinformation in toxic interactions online. We outline the limitations of our approach in this section.

Unreliable Information. One of the limitations of our approach is our use of hyperlinks to determine the presence of unreliable/factually inaccurate information. As we examined much of Reddit's 1.6B comments, we were unable to take a comment-by-comment-based approach to understand the spread of and the levels of unreliable news. As a result, our approach inevitably missed some subtleties of unreliable information across subreddits. However, as found in several past works [65, 75, 139, 158], examining unreliable information from a domain-based perspective enables researchers to track readily identifiable and questionable information across different platforms and is a reliable way of understanding the presence of unreliable information in large communities or websites (*e.g.*, subreddits).

We further note that our approach relies on the presence of largely US-based domains on given subreddits and only measures English unreliable information and US-based US-partisanship. As a result, we cannot simply apply our results to non-English subreddits. However, we note that while our work centers on US-based political environments, as found in prior works, highly political environments across different cultures often utilize unreliable information and often share many of the same characteristics as US ones [65, 82]. We leave the full investigation of these settings on Reddit to future work.

Measuring Toxicity. Another limitation of our approach, given our use of the Perspective API to estimate toxicity based on the comment load, is that it is limited to relatively 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 100 comments. As such, our results are skewed to more active subreddits and users. At the same time, these subreddits and users make up a large percentage of users' experiences on Reddit.

Confounds, Correlation, and Causation. We lastly acknowledge that while we account for many user-level and subreddit-level features, there may be other hidden confounders. For example, while we attempted to remove automated accounts from much of our analysis by removing accounts that were labeled as "bot" accounts, due to the rapid rise of AI, within Reddit as a whole, there could still be automated accounts. We note that we conducted this analysis for data in 2020 and 2021, before the release of ChatGPT, however. We further emphasize that while we work to account for confounders, the results we present describe the correlation between misinformation, political polarization, and toxicity; we cannot ascribe causation. Similarly, we note that while our approach does attempt to take into account the partisanship and politicalness of different subreddits and users, we currently do not seek to fully understand *qualitatively* why unreliable domain submissions to Reddit are correlated with increased toxicity (*i.e.*, more "controversial" topics broached, the information presented just being false, users arguing about the post not being taken down). However, our results that illustrate an increased toxicity in posts that include hyperlinks from unreliable websites and false information do align with a large literature of similar results [11–13], some of which have found causal results.

Use of 2020-2021 Reddit Data. Within this study, we utilize Reddit data posted between January 2020 to June 2021, which at the time of collection was still available through Pushshift [14]. While we note that this is relatively older data, this is largely due to public restrictions on the use of the Reddit API [35]. We note our dataset, while older, contains over 1.6B comments across 57.2K subreddits and 15.5M users, bolstering the generalizability of our work, particularly in comparison to similar studies [169]. We note that our work does not focus on the temporal nature of Reddit but rather on the particular political and toxicity norms of particular subreddits and how users interact with the presence of unreliable and reliable information from third-party websites.

7 Discussion

In this work, we examined the relationship between unreliable information, political partisanship, user engagement, and toxicity across and within both political and non-political subreddits. Using previously published lists of unreliable and reliable websites, we find that on Reddit, we find that comments posted in response to submissions with hyperlinks to unreliable news websites have 25% more toxic comments overall (an average of 32% more within individual subreddits). Utilizing a zero-inflated negative binomial model to model engagement with unreliable versus reliable information sources, we observe that subreddit toxicity is a major predictor of whether unreliable domain submissions receive comments. This contrasts with reliable domain submissions, where toxicity plays a more muted role. Finally, examining how partisanship and unreliable information affect the increase in toxic interactions, we find, confirming with an Exponential Random Graph Model (ERGM), that articles from unreliable news outlets, compared to those from reliable news outlets, correlate with increased toxicity among users of different political leanings (*i.e.*, affective polarization).

7.1 Unreliable Information, Polarization, and Toxicity

Our work shows that while unreliable websites have much less of a presence on Reddit compared to reliable websites (633.6K posts/601 submissions per domain vs 7.55M posts/2010.4 submissions per domain), unreliable news websites play a significant role on the platform. In addition to misleading users, unreliable information's effect on the discourse on these subreddits can often be pernicious, with articles from websites known to promote unreliable news increasing inter-political strife. Indeed, as was seen in Table 2 and was found in our unreliable domain submission dyads, unreliable domain submissions are associated with increased toxicity, particularly among users of different partisanship alignments. This largely accords with the work of Dicicco et al. [37] that showed that users who comment on YouTube videos promoting COVID-19 conspiracy theories often utilize toxic and vulgar language. Our paper results bolster this work, showing that increased unreliable domains correlate with increased incivility on Reddit. This largely goes to promote and affirm the view that unreliable news/misinformation does have a relationship [37, 111] with user toxicity and is not uncorrelated with toxicity [31, 121].

In our conversational dyads, we further find that across much of Reddit, unreliable websites are correlated with more insular and politically one-sided conversations, while reliable domains are correlated with increased discussions between users of different political ideologies (with both increasing inter-political toxicity). Community norms for particular environments appear to further affect how users engage with different materials. As found with our zero-inflated negative binomial model, subreddit toxicity norms are also predictive of user engagement with unreliable news articles. The more toxic/uncivil a given environment, the more likely at least one person will engage with misinformation or unreliable sources. However, simultaneously, in more toxic environments, where these posts most commonly appear, these same posts are less likely to gain extensive engagement and a large number of comments. This appears to reflect that unreliable news websites often utilize "clickbait" titles that induce readers to initially comment, but then cause the reader to not often thoroughly engage with the material [26, 120]. In contrast, in less toxic environments where these factually unreliable Reddit submissions more rarely appear, if they do gain traction (*e.g.*, at least one comment), they are more likely to gain more comments.

Reddit Idiosyncrasies: Comparison to Other Platforms. In this work, across 18 months of data and over 1.6B comments, we find that within the stable decentralized online communities present on Reddit, there was increased toxicity whenever users interacted with unreliable website submissions. This decentralization and stability further allowed for substantial variation and marked differentiation in partisanship, toxicity, and tolerance to unreliable information. As a result, within different corners of Reddit, as seen in Section 4, there were very different relationships between the subreddit's partisan and toxicity characteristics and the presence of unreliable submissions [125]. In contrast, platforms like X, Facebook, or YouTube, which employ centralized algorithmic moderation and that do not have structured communities with stated particular or developed values, often exert a more uniform influence on content visibility and enforcement; as a result, there are likely to not be as many particular public communities on these platforms that would enable these toxicity norms and tolerances to form. As a result, this may explain mixed results that have been found on platforms like X and YouTube [31, 111, 112, 121] as different users with different values and norms insert and exit particular conversations on the platforms. For example, one community that may emerge on X may focus on sharing particular types of news [72], but that community may change rapidly over time and never establish particular norms. Indeed, as found by Hanley et al. [64], often within conversations surrounding individual topics, the compositions of users involved can rapidly change on X.

7.2 Implications for the Reddit Platform

Our work indicates that unreliable website submissions correlate with increased overall toxicity of conversations on Reddit, particularly between users of different partisan affiliations. We note that this increased rancor persists despite individual subreddits moderating unreliable domain submissions more heavily compared to reliable domain submissions. Given the lower prevalence of unreliable sources throughout Reddit compared to reliable sources and the decreased toxicity of conversations with moderator involvement, a potential solution to decrease toxicity may be for Reddit admins (who are not already doing so) to engage more thoroughly or to flag submissions that contain hyperlinks to known unreliable and specious websites. However, as argued by Bozrath et al. [18], different approaches for moderating this content in different subreddits, however, will be necessary. Some larger subreddits already take a machine-learning approach to remove misinformation [84] while others have taken a manual approach that relies on crowd wisdom or individual moderator involvement [81, 87, 137]. As found by Kim et al. [90], there is a differential impact when communities fact-check information versus individuals; when individual or fact-checking organizations label posts as false, users often retreat into information echo chambers compared to when they are fact-checked by their community or through collective means (*e.g.*, community notes). We argue that Reddit is uniquely positioned to take advantage of this behavior, given that users' communities already upvote and downvote information. We argue that this learned behavior could be oriented toward assigning community notes to posts flagged by moderators or admins. Given that Reddit has shown a penchant towards being proactive on propaganda (the website removed links to Russian state-based propaganda in the wake of the Russo-Ukrainian War [144] and has previously taken steps to remove highly toxic material and subreddits [143]), we recommend that Reddit itself also take more proactive steps to alert users to unreliable information and use community notes as enforcement mechanism. We note that this approach has become increasingly popular as platforms like Facebook, Instagram, and Threads adopt it instead of using third-party fact-checkers [101]. Further, as again found by Bozrath et al. [18] moderating one type of misinformation or unreliable source can be similar to moderating other types, allowing Reddit to take a generalized approach to alert subreddits to the presence of unreliable news and propaganda.

Furthermore, as found by [93], use of these types of fact-checks are effective in "reducing the perceived accuracy of false headlines."

We note, as found in this study and in recent prior work [20], unreliable information is largely concentrated within a few communities on Reddit and is largely dwarfed by content from reliable and factual news organizations. While our study does indicate that unreliable information does lead to increased toxicity, we note that its overall prevalence compared to reliable news sources *is* low. We further argue that targeted campaigns to remove bad actors/communities on Reddit could be effective [74, 133]. As argued by Ribeiro et al. [128], deplatforming these actors, while leading to less engagement on Reddit specifically, also could lead to less engagement for these actors across the web. Depending on the severity of unreliable information, we argue that Reddit should take action per this tradeoff.

Political Echo-Chambers, Politics Discussions, and Reliable News on Reddit. Similar to past work, we find that most interactions take place among users of the same political orientation [43]. Reddit specifically creates communities for like-minded people, and as a result, most interactions (both toxic and non-toxic interactions) on the platform are between people of the same political orientation. However, most interestingly, in the comments of submissions with hyperlinks to reliable news sources, the rate of inter-partisan interactions slightly increases compared to interaction across Reddit. This is in contrast to unreliable domain submissions, where the rates of interaction between users of different partisan affiliations decrease. We argue that if Reddit, as a whole, desires to lower levels of political incivility, fewer echo-chambers, and less toxicity on its platform, taking a more proactive approach to policing questionable sources could help alleviate these issues. As found by Gallacher et al. [52], toxic online interactions between political groups often lead to offline real-world political violence. Given that unreliable news appears to be correlated with and reinforces toxic interactions between different political groups, this highlights the need to research its effects and curtail its spread.

Sub-Standards/Community Norms. We have found throughout this work that subreddits interact with reliable and unreliable sources differently. For example, while more toxic subreddits are more likely to interact with unreliable information sources, the more toxic a subreddit, the less likely the reliable submissions are to get any comments. We thus find that often complex relationships exist between different types of subreddits and their interactions with different types of posts. There is no one-size-fits-all approach to understanding user engagement and toxicity on Reddit [138, 174]. We thus argue that a subreddit/community-based approach that takes into account the community norms of the community must be taken when trying to understand the information flows within it [45]. Similarly, in attempting to prevent engagement with unreliable news on particular subreddits, understanding their toxicity norms, their political ideology, and who is posting the article within the subreddit is key [174]. For example, as found by Zhan et al. [174], different communities responded and engaged with COVID-19 misinformation in widely divergent manners. We thus argue that approaches that attempt to understand how users engage with unreliable information (particularly on Reddit), *must* take into account the particular nuances of that community.

8 Conclusion

Unreliable information persists across many types of subreddits, and its spread appears to be influenced by the nature of the community in which it is posted. Factually incorrect content is more likely to gain traction in toxic or uncivil environments. Communities with high levels of unreliable news tend to be more right-leaning and politically insular, with interactions occurring primarily among users with similar political orientations. Our ERGM analysis further shows that when users commenting on unreliable information are more politically dissimilar, they are also more likely to engage in toxic or uncivil exchanges compared to those commenting on reliable information. This study—one of the first to examine the relationship between unreliable news, toxicity, and political ideology at scale—highlights the need to fully understand the impact of unreliable information in order to combat political polarization and online toxicity. Unreliable news not only misleads people but may also amplify political divisions and be correlated with more toxic online environments.

Acknowledgments

This work was supported in part by the NSF Graduate Fellowship DGE-1656518, a Meta Ph.D. Fellowship, and a Sloan Research Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation or other funding agencies.

References

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

[2] 2022. Metrics For Reddit - Complete List Of Subreddits - Updated Weekly. https://frontpagemetrics.com/list-all-subreddits

[3] 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*.

[4] 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.

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

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

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

[8] Hind Almerekhi, Supervised by Bernard J Jansen, and co-supervised by Haewoon Kwak. 2020. Investigating toxicity across multiple Reddit communities, users, and moderators. In *Companion proceedings of the web conference 2020*. 294–298.

[9] Tawfiq Ammari, Sarita Schoenebeck, and Daniel Romero. 2019. Self-declared throwaway accounts on Reddit: How platform affordances and shared norms enable parenting disclosure and support. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[10] Jisun An, Haewoon Kwak, Oliver Posegga, and Andreas Jungherr. 2019. Political discussions in homogeneous and cross-cutting communication spaces. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 68–79.

[11] 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.

[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] 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.

[15] 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.

[16] Robert M Bond and Matthew D Sweitzer. 2022. Political homophily in a large-scale online communication network. *Communication Research* 49, 1 (2022), 93–115.

[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] Lia Bozarth, Jane Im, Christopher Quarles, and Ceren Budak. 2023. Wisdom of Two Crowds: Misinformation Moderation on Reddit and How to Improve this Process—A Case Study of COVID-19. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–33.

[19] Lia Bozarth, Aparajita Saraf, and Ceren Budak. 2020. Higher ground? How groundtruth labeling impacts our understanding of fake news about the 2016 US presidential nominees. In *Proceedings of the International AAAI*

Conference on Web and Social Media, Vol. 14. 48–59.
[20] Ceren Budak, Brendan Nyhan, David M Rothschild, Emily Thorson, and Duncan J Watts. 2024. Misunderstanding the harms of online misinformation. *Nature* 630, 8015 (2024), 45–53.
[21] Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. 2022. Finding Qs: Profiling QAnon Supporters on Parler. https://doi.org/10.48550/ARXIV.2205.08834
[22] 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.
[23] Global Engagement Center. 2020. Pillars of Russia's disinformation and propaganda ecosystem.
[24] Pew Research Center. 2017. The partisan divide on political values grows even wider. *Pew Research Center* (2017).
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] 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.
[30] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.
[31] 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.
[32] 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.
[33] Dana Cuomo and Natalie Dolci. 2019. Gender-Based Violence and Technology-Enabled Coercive Control in Seattle: Challenges & Opportunities.
[34] Alina Darmstadt, Mick Prinz, and Oliver Saal. 2019. The murder of Keira: misinformation and hate speech as far-right online strategies. (2019).
[35] Paresh Dave. 2023. Reddit Is Already on the Rebound. https://www.wired.com/story/reddit-is-already-on-the-rebound/.
[36] 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.
[37] Karen DiCicco, Nahiyan B Noor, Niloofar Yousefi, Maryam Maleki, Billy Spann, and Nitin Agarwal. 2020. Toxicity and Networks of COVID-19 discourse communities: a tale of two social media platforms. *Proceedings http://ceur-ws. org ISSN* 1613 (2020), 0073.
[38] 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.
[39] James N Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2021. Affective polarization, local contexts and public opinion in America. *Nature human behaviour* 5, 1 (2021), 28–38.
[40] Maeve Duggan. 2017. Online Harassment 2017 | Pew Research Center. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/
[41] Alexander Dyck and Luigi Zingales. 2003. *The media and asset prices*. Technical Report. Working Paper, Harvard Business School.
[42] 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.
[43] Alexandros Efstratiou, Jeremy Blackburn, Tristan Caulfield, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. 2023. Non-polar opposites: analyzing the relationship between echo chambers and hostile intergroup interactions on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 197–208.
[44] Facebook. 2021. Transparency center. https://transparency.fb.com/policies/community-standards/bullying-harassment/datz. Accessed: 2021-10-08.

[45] 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*.

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

[47] 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.

[48] 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*.

[49] 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.

[50] 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.

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

[52] 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.

[53] 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.

[54] 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.

[55] 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.

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

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

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

[59] Doris Appel Graber, Denis McQuail, and Pippa Norris. 1998. *The politics of news: The news of politics*. CQ press Washington, DC.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] Hans WA Hanley and Zakir Durumeric. 2023. Twits, toxic tweets, and tribal tendencies: Trends in politically polarized posts on twitter. *arXiv preprint arXiv:2307.10349* (2023).

[65] 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.

[66] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 2023. " A Special Operation": A Quantitative Approach to Dissecting and Comparing Different Media Ecosystems' Coverage of the Russo-Ukrainian War. In *Proceedings of the International AAAI Conference on Web and social media*, Vol. 17. 339–350.

[67] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 2023. A golden age: conspiracy theories' relationship with misinformation outlets, news media, and the wider internet. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–33.

[68] 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. In *Proceedings of the international AAAI conference on web and social media*, Vol. 17. 327–338.

[69] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 2024. Specious sites: Tracking the spread and sway of spurious news stories at scale. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1609–1627.
[70] Hans WA Hanley, Emily Okabe, and Zakir Durumeric. 2025. Tracking the Takes and Trajectories of English-Language News Narratives across Trustworthy and Worrisome Websites. In *34th USENIX Security Symposium (USENIX Security 25)*.
[71] Gordon Heltzel and Kristin Laurin. 2020. Polarization in America: Two possible futures. *Current Opinion in Behavioral Sciences* 34 (2020), 179–184.
[72] AmaÇ HerdaĞdelen, Wenyun Zuo, Alexander Gard-Murray, and Yaneer Bar-Yam. 2013. An exploration of social identity: The geography and politics of news-sharing communities in twitter. *Complexity* 19, 2 (2013), 10–20.
[73] Marc J Hetherington. 2008. Turned off or turned on? How polarization affects political engagement. *Red and blue nation* 2 (2008), 1–33.
[74] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–24.
[75] 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*.
[76] 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).
[77] 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.
[78] 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.
[79] 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.
[80] 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.
[81] Sohyeon Hwang and Jeremy D Foote. 2021. Why do people participate in small online communities? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.
[82] 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.
[83] 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.
[84] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.
[85] 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.
[86] Google Jigsaw. 2022. Google Jigsaw. Perspective API. https://www.perspectiveapi.com/home.
[87] Ridley Jones, Lucas Colusso, Katharina Reinecke, and Gary Hsieh. 2019. r/science: Challenges and opportunities in online science communication. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.
[88] John T Jost, Sander van der Linden, Costas Panagopoulos, and Curtis D Hardin. 2018. Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation. *Current opinion in psychology* 23 (2018), 77–83.
[89] Julia Kamin. 2019. *Social Media and Information Polarization: Amplifying Echoes or Extremes?* Ph. D. Dissertation.
[90] Junsol Kim, Zhao Wang, Haohan Shi, Hsin-Keng Ling, and James Evans. 2025. Differential impact from individual versus collective misinformation tagging on the diversity of Twitter (X) information engagement and mobility. *Nature Communications* 16, 1 (2025), 973.
[91] 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.
[92] 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.

[93] Jan Kirchner and Christian Reuter. 2020. Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness. *Proceedings of the ACM on Human-computer Interaction* 4, CSCW2 (2020), 1–27.

[94] Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 865–878.

[95] Deepak Kumar, Jeff Hancock, Kurt Thomas, and Zakir Durumeric. 2023. Understanding the behaviors of toxic accounts on reddit. In *Proceedings of the ACM Web Conference 2023*. 2797–2807.

[96] 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.

[97] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference*. 933–943.

[98] 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).

[99] 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.

[100] Alex Leavitt. 2015. " This is a Throwaway Account" Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 317–327.

[101] Anne Marie D. Lee. 2025. What is Community Notes, and how will it work on Facebook and Instagram? https://www.cbsnews.com/news/what-is-community-notes-twitter-x-facebook-instagram/

[102] 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.

[103] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. All that's happening behind the scenes: Putting the spotlight on volunteer moderator labor in Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 584–595.

[104] 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.

[105] Rijul Magu, Nivedhitha Mathan Kumar, Yihe Liu, Xander Koo, Diyi Yang, and Amy Bruckman. 2024. Understanding Online Discussion Across Difference: Insights from Gun Discourse on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–28.

[106] 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.

[107] Michalis Mamakos and Eli J Finkel. 2023. The social media discourse of engaged partisans is toxic even when politics are irrelevant. *PNAS nexus* 2, 10 (2023), pgad325.

[108] Abdulla Mamun and Sudhir Paul. 2023. Model selection in generalized Linear models. *Symmetry* 15, 10 (2023), 1905.

[109] 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.

[110] 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.

[111] Mohsen Mosleh, Rocky Cole, and David G Rand. 2024. Misinformation and harmful language are interconnected, rather than distinct, challenges. *PNAS nexus* 3, 3 (2024), pgae111.

[112] Mohsen Mosleh and David G Rand. 2022. Measuring exposure to misinformation from political elites on Twitter. *Nature Communications* 13, 1 (2022), 7144.

[113] Matti Nelimarkka, Salla-Maaria Laaksonen, and Bryan Semaan. 2018. Social media is polarized, social media is polarized: towards a new design agenda for mitigating polarization. In *Proceedings of the 2018 designing interactive systems conference*. 957–970.

[114] 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.

[115] Henna Paakki, Heidi Vepsäläinen, Antti Salovaara, and Bushra Zafar. 2024. Detecting covert disruptive behavior in online interaction by analyzing conversational features and norm violations. *ACM Transactions on Computer-Human Interaction* 31, 2 (2024), 1–43.

[116] 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.

[117] Se Jung Park, Yon Soo Lim, and Han Woo Park. 2015. Comparing Twitter and YouTube networks in information diffusion: The case of the "Occupy Wall Street" movement. *Technological forecasting and social change* 95 (2015), 208–217.

[118] 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.

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

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

[121] Alessandro Quattrociocchi, Gabriele Etta, Michele Avalle, Matteo Cinelli, and Walter Quattrociocchi. 2022. Reliability of news and toxicity in twitter conversations. In *International Conference on Social Informatics*. Springer, 245–256.

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

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

[124] Ashwin Rajadesingan, Ceren Budak, and Paul Resnick. 2021. Political discussion is abundant in non-political subreddits (and less toxic). In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 525–536.

[125] 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.

[126] Jonas Ranstam and Jonathan A Cook. 2018. LASSO regression. *Journal of British Surgery* 105, 10 (2018), 1348–1348.

[127] Steve Rathje, Jay J Van Bavel, and Sander Van Der Linden. 2021. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences* 118, 26 (2021), e2024292118.

[128] Manoel Horta Ribeiro, Shagun Jhaver, Marie Reignier-Tayar, Robert West, et al. 2024. Deplatforming norm-violating influencers on social media reduces overall online attention toward them. *arXiv preprint arXiv:2401.01253* (2024).

[129] 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.

[130] 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.

[131] 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.

[132] Dana Rotman, Jennifer Golbeck, and Jennifer Preece. 2009. The community is where the rapport is–on sense and structure in the youtube community. In *Proceedings of the fourth international conference on Communities and technologies*. 41–50.

[133] Giuseppe Russo, Luca Verginer, Manoel Horta Ribeiro, and Giona Casiraghi. 2023. Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning. In *Proceedings of the international AAAI conference on web and social media*, Vol. 17. 742–753.

[134] 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.

[135] 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.

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

[137] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The social roles of bots: evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–29.

[138] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2022. Metaphors in moderation. *New Media & Society* 24, 3 (2022), 621–640.

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

[140] 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.

[141] 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.

[142] 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.

[143] Todd Spangler. 2020. Reddit Finally Bans Hate Speech, Removes 2,000 Racist and Violent Forums Including The_Donald. https://variety.com/2020/digital/news/reddit-bans-hate-speech-groups-removes-2000-subreddits-donald-trump-1234692898/.

[144] Todd Spangler. 2022. Reddit Bans Links to Russian State Media Across Entire Site. https://variety.com/2022/digital/news/reddit-bans-links-to-russian-state-media-1235195612/.

[145] 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*.

[146] Jennifer Stromer-Galley. 2003. Diversity of political conversation on the Internet: Users' perspectives. *Journal of Computer-Mediated Communication* 8, 3 (2003), JCMC836.

[147] Cass R Sunstein. 2018. Is social media good or bad for democracy. *SUR-Int'l J. on Hum Rts.* 27 (2018), 83.

[148] 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.

[149] 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.

[150] Amaury Trujillo and Stefano Cresci. 2022. Make reddit great again: assessing community effects of moderation interventions on r/the_donald. *Proceedings of the ACM on Human-computer Interaction* 6, CSCW2 (2022), 1–28.

[151] Milo Trujillo, Maurício Gruppi, Cody Buntain, and Benjamin D Horne. 2020. What is bitchute? characterizing the. In *Proceedings of the 31st ACM conference on hypertext and social media*. 139–140.

[152] 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).

[153] 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.

[154] Theresa Ullmann, Georg Heinze, Lorena Hafermann, Christine Schilhart-Wallisch, Daniela Dunkler, and TG2 of the STRATOS initiative. 2024. Evaluating variable selection methods for multivariable regression models: A simulation study protocol. *Plos one* 19, 8 (2024), e0308543.

[155] Hilde Van den Bulck and Aaron Hyzen. 2020. Of lizards and ideological entrepreneurs: Alex Jones and Infowars in the relationship between populist nationalism and the post-global media ecology. *International communication gazette* 82, 1 (2020), 42–59.

[156] 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.

[157] 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.

[158] 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).

[159] Isaac Waller and Ashton Anderson. 2019. Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. In *The World Wide Web Conference*. 1954–1964.

[160] Isaac Waller and Ashton Anderson. 2021. Quantifying social organization and political polarization in online platforms. *Nature* 600, 7888 (2021), 264–268.

[161] 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*.

[162] 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.

[163] Galen Weld, Maria Glenski, and Tim Althoff. 2021. Political bias and factualness in news sharing across more than 100,000 online communities. In *Proceedings of the international AAAI conference on web and social media*, Vol. 15. 796–807.

[164] 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.

[165] Tom Wilson and Kate Starbird. 2020. Cross-platform disinformation campaigns: Lessons learned and next steps. *Harvard Kennedy School Misinformation Review* (2020).

[166] 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.

[167] 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.

[168] 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.

[169] Yulin Yu, Julie Jiang, and Paramveer S Dhillon. 2024. Characterizing the Structure of Online Conversations Across Reddit. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–23.

[170] Dave Van Zandt. 2025. Media-Bias/Fact-Check. https://mediabiasfactcheck.com/yahoo-news/.

[171] 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.

[172] 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.

[173] 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.

[174] Jason Shuo Zhang, Brian Keegan, Qin Lv, and Chenhao Tan. 2021. Understanding the diverging user trajectories in highly-related online communities during the COVID-19 pandemic. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 888–899.

A Embeddings Hyperparameter Optimization

Variable	Values Considered
Embedding Size	100, 150, 200
Number of negative examples	30, 35, 40, 45
Down-Sampling threshold;	0.0025 0.005, 0.0075, 0.01
The starting learning rate	0.15, 0.18, 0.21

Table 7. We optimize our community and user embeddings.

B Goodness of Fit For EGRMs

Unreliable Domain Interactions	**MC p-value**	**Reliable Domain Interactions**	**MC p-value**
Intercept/Edges	0.76	Intercept/Edges	0.86
User Partisanship Differences	0.84	User Partisanship Differences	0.86
User Toxicity	0.48	User Toxicity	0.90
Shared Subreddits Comments	0.90	Shared Subreddits Comments	0.68
Reciprocity	0.84	Reciprocity	0.78

Table 8. **Goodness-of-fit via Monte Carlo p-values for ERGMs modeling interactions with unreliable and reliable news domains.** The p-values indicate how well the specified ERGM replicates key network features under a goodness-of-fit test. Low values (typically < 0.05) indicate a lack of fit. We note that these terms summarize how well the entire fitted model, which includes all terms, reproduces specific network-level features (e.g., degree distribution, shared partners, *etc...*) that are influenced by each term.

Received October 2024; revised April 2025; accepted August 2025